MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

④

AD-A189 035

Mental Models of Mechanical Systems:

Individual Differences in Qualitative

and Quantitative Reasoning.

Mary Hegarty

Marcel Adam Just

Ian R. Morrison

DEPARTMENT
of
PSYCHOLOGY

DTIC
S ELECTE D
JAN 2 2 1988
D

Carnegie-Mellon University

88 1 13 0

Mental Models of Mechanical Systems:

Individual Differences in Qualitative

and Quantitative Reasoning.

Mary Hegarty

Marcel Adam Just

Ian R. Morrison

To be published in <u>Cognitive Psychology</u>

DTIC
ELECTE
JAN 2 2 1988
S
D

REPORT DOCUMENTATION PAGE   A187035

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; Distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ONR87-8	5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION Carnegie-Mellon University	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Personnel & Training Research Programs Office of Naval Research (Code 1142PT)
6c. ADDRESS (City, State, and ZIP Code) Department of Psychology Pittsburgh, PA 15213		7b. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, VA 22217

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-85-K-0584
8c. ADDRESS (City, State, and ZIP Code)		10. SOURCE OF FUNDING NUMBERS

8c. ADDRESS (City, State, and ZIP Code)	PROGRAM ELEMENT NO. 61153N	PROJECT NO. RR04206	TASK NO. RR04206-0C	WORK UNIT ACCESSION NO. NR702-014

11. TITLE (Include Security Classification)
Mental Models of Mechanical Systems: Individual Differences in Qualitative and Quantitative Reasoning

12. PERSONAL AUTHOR(S)
Mary Hegarty, Marcel Adam Just, & Ian R. Morrison

13a. TYPE OF REPORT Technical Report	13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Year, Month, Day) 1987, Dec	15. PAGE COUNT 34

16. SUPPLEMENTARY NOTATION

17.	COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Mental models, Individual differences,
05	09		Mechanical ability, Production system simulation

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

People who understand mechanical systems can infer the principles of operation of an unfamiliar device from their knowledge of the devices's components and their mechanical interactions. Individuals vary considerably in their ability to make this type of inference. This paper describes studies of performance in psychometric tests of mechanical ability. Based on subjects' retrospective protocols and response patterns, it was possible to identify rules of mechanical reasoning that accounted for the performance of subjects of different levels of mechanical ability. The rules are explicitly stated in a simulation model which demonstrates the sufficiency of the rules by producing the kinds of responses observed in the subjects. Three abilities are proposed as the sources of individual differences in performance: (1) ability to correctly identify which attributes of a system are relevant to its mechanical function, (2) ability to use rules consistently, and (3) ability to quantitatively combine information about two or more relevant attributes. Keywords

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Susan Chipman	22b. TELEPHONE (Include Area Code) (202) 696-4318	22c. OFFICE SYMBOL ONR1142PT

DD FORM 1473, 34 MAR 83 APR edition may be used until exhausted.   SECURITY CLASSIFICATION OF THIS PAGE
All other editions are obsolete.

Abstract

People who understand mechanical systems can infer the principles of operation of an unfamiliar device from their knowledge of the device's components and their mechanical interactions. Individuals vary considerably in their ability to make this type of inference. This paper describes studies of performance in psychometric tests of mechanical ability. Based on subjects' retrospective protocols and response patterns, it was possible to identify rules of mechanical reasoning that accounted for the performance of subjects of different levels of mechanical ability. The rules are explicitly stated in a simulation model which demonstrates the sufficiency of the rules by producing the kinds of responses observed in the subjects. Three abilities are proposed as the sources of individual differences in performance: (1) ability to correctly identify which attributes of a system are relevant to its mechanical function, (2) ability to use rules consistently, and (3) ability to quantitatively combine information about two or more relevant attributes.

According to Greek legend, Archimedes was once asked by Hieron, the king of Syracuse, to demonstrate how his theory of mechanics would allow a very heavy weight to be lifted by a very small force. Archimedes responded by constructing a pulley system that permitted Hieron to lift a heavily laden ship with the force of his own arm. Archimedes' understanding of pulley systems surpasses the mechanical ability of most educated people today, and highlights the fact that there are vast differences among individuals in this ability. We generally associate mechanical ability with a person's understanding of how machines work, the ability to build a machine out of its elementary components, and the ability to determine why a machine is not working correctly. To understand a machine in this way, a person has to be able to identify the main components of the machine, know which properties of these components are relevant to their function in the system, and also understand how these components interact to accomplish the machine's function. This paper explores the nature of mechanical ability, and provides an account of individual differences in mechanical ability.

Our approach to studying mechanical ability has been to determine what kinds of rules people of different ability use to relate the attributes of a mechanical system to its function. The research examines which attributes of a mechanical system people consider relevant to its function, their rules relating the attributes to the function, their preferences among different rules, and their methods for combining rules pertaining to different attributes. The methodology includes an analysis of verbal protocols as well as an analysis of the response patterns obtained during the performance on test items, and these analyses permit us to infer which rules are used by people of different ability. The resulting models of high and low ability subjects are instantiated as two computer simulation models, whose performance on the test items produces patterns resembling those of human subjects.

Previous studies of mechanical ability used the psychometric approach, which involves measuring the correlations between performance on tests of mechanical ability and other basic abilities. Psychometric analyses suggested that there were several components of mechanical ability, such as general reasoning ability, as well as knowledge acquired through experience with machines (Cronbach, 1984). Cast in this light, the study of individual differences in mechanical ability is interesting for several reasons. First, if mechanical reasoning reflects knowledge, we can use mechanical reasoning tests to study different levels of understanding of machines which characterize high and low ability people. Second, it suggests that mechanical ability may not be a static trait but may develop with experience, so that the study of individual differences may also have implications for learning and development. Third, it suggests that mechanical ability is an instance of reasoning about a particular domain and its underlying principles of operation, so that the characteristics of mechanical reasoning may also generalize to reasoning in other domains such as understanding a biological system or a social organization.

The rules that a subject uses to relate the attributes of a machine to its function reflect his level of understanding of the machine. There are several different levels of understanding of machines that a person might have acquired through experience. People who use some simple machines in everyday life may understand how these machines work but be unable to extrapolate this knowledge to understanding an unfamiliar machine. Alternatively, people might have abstracted from their experience some general principles of machines. One such principle might be the relation between the attributes of a machine and its mechanical advantage (the amount by which the machine magnifies input force). Such principles can be used in understanding an unfamiliar machine. Finally people who know some formal physics might be able to analyze the balance of forces in an unfamiliar system and calculate a precise value for the mechanical advantage of the system.

Different levels of understanding of machines can be illustrated in the history of scientific understanding of a simple machine such as the pulley system. According to the Encyclopedia Britannica, pulleys were first used in about the 8th century B.C., considerably later than other simple machines, such as the lever and wedge. Like all the simple machines, pulleys were used long before the mathematical relationships between loads and displacements in these machines were formally described by Archimedes (3rd century B.C.). The analysis of pulley systems in terms of the balance of force in the systems depends on principles of Newtonian physics, formalized about 2000 years later. Thus, it is obviously possible to have practical understanding of pulley systems without understanding the physics principles that underlie their operation.

Different levels of understanding of machines have previously been studied as expert novice differences. Studies of novice understanding of machines suggests that as a result of everyday experience with machines, people develop intuitive physical laws which are typically qualitative, situation-specific, and involve misconceptions (Clement, 1983; diSessa, 1983; White, 1983). They can be contrasted with the laws of physics, known by experts, which are quantitative and are consistent in explaining a wide range of physical phenomena. A special case of an expert-novice difference is provided by the contrast between children of different ages. In studying the development of children's understanding of the balance beam, Siegler (1978) found that very young children make errors by failing to take some important attribute of the machine into consideration, while somewhat older children take all relevant attributes into account but are unable to combine information about two relevant attributes. The types of cognitive differences that separate experts from novices, and older children from younger children, may represent a systematic progression of mental models that characterize individual differences in mechanical ability.

The Experimental Task. We studied individual differences in the type of task used in psychometric tests of mechanical ability. The early psychometric tests that were designed to measure mechanical ability were performance tests in which the subjects had to manipulate real, three dimensional objects (Cox, 1928). However, it was later discovered that the actual physical manipulation of an object was not essential to the validity of the test (Stenquist, 1922; Smith, 1964). Paper-and-pencil tests of mechanical ability have been found to be comparably predictive of performance in a number of technical fields such as machine assembly, mechanical repair, electrical work and vehicle operation (Bennett, 1969; Ghiselli, 1955; Vernon & Parry, 1949). It is such paper-and-pencil tests that we analyze in the current research. Our approach has been to analyze performance on items like those in the psychometric test itself, on the assumption that the processes that are sources of individual differences in test performance are also sources of individual differences in mechanical ability in the real world.

The research takes as its starting point the Bennett Mechanical Comprehension Test (Bennett, 1969), one of the most widely-used tests of mechanical ability (Bechtoldt, 1972). According to the manual for this test, its objective is to "measure the ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations." The Bennett items require qualitative rather than quantitative reasoning, such as being able to compare two depicted mechanical systems in terms of the relative amount of input force they require to achieve their mechanical function, rather than being able to compute a precise value for the mechanical advantage of a particular system.

Although the Bennett test itself contains items pertaining to many aspects of mechanics (such as fluid and thermal dynamics, levers, gears, and pulleys), our experiments focused exclusively on pulley problems. The focus on pulleys permitted us to construct a large number of pulley problems that systematically varied the number and type of

attributes that distinguished the two systems depicted in each problem. Pulleys are prototypical of machines, containing a set of physically interacting components that allow a user to multiply force at the expense of distance. And the Bennett type of pulley problems were at an appropriate difficulty level for our subjects, allowing measurement of a range of individual differences in performance. In a pilot study, the mean proportion of pulley problems solved correctly (.62) was lower and the variance greater (S.D. = .26) than problems involving other simple machines, such as levers and gears. Restricting the experiments to pulley problems does not compromise the generality of the research, since previous analyses of the Bennett test (Cronbach, 1984) and our own pilot study have shown that separate scores for different types of items are highly correlated. Thus our examination of the mechanical ability that deals with pulleys should apply to reasoning about other types of mechanical systems.

A pulley problem of the type used in the Bennett test and in our study is shown in Figure 1. In this problem, the subject is asked to decide which of two pulley systems will require more force to lift a weight. The test instructions state that the pulleys in the systems are weightless and frictionless. To understand the problem, a subject must know how the forces balance in the two depicted pulley systems. If the system is in equilibrium, the force is equal throughout the rope, and the sum of the upward forces at any point in the system is equal to the sum of the downward forces. If the person using pulley system B exerts a unit force on the pull rope, there will be a force equal to two units acting on the weight because there are two rope strands pulling up on the weight, one on either side of the movable pulley. We will refer to the amount of force required to lift a weight with a pulley system as the effort. The ratio of the weight to be lifted to the effort is the mechanical advantage of the pulley system. Thus system B has a mechanical advantage of two. For example, the person could support a 20 lb weight by exerting a force of 10 lbs on the rope. In the case of pulley system A, if the person exerts a unit force on the pull rope, there will be two unit forces supporting the upper movable pulley. The force on the lower rope will be twice the value of the input force and there are two strands of this rope pulling up on the weight, so there will be a total of four unit forces pulling up on the weight and the person could support a 20 lb weight by exerting a force of 5 lbs on the rope. System A has a mechanical advantage of four. Pulley system B therefore has less mechanical advantage than pulley system A and so the person has to pull with more force when using pulley system B.
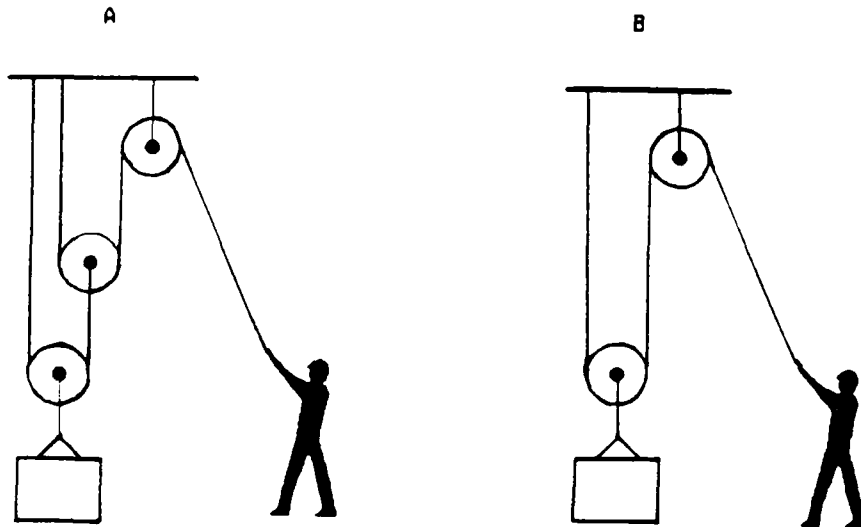

Insert Figure 1 about here.

People can sometimes successfully solve the problems in our study by comparing the depicted pulley systems on the basis of a number of visible attributes such as the number of load-bearing ropes and the number of pulleys. Such attributes covary with mechanical advantage in the sense that pulley systems with higher values on these attributes tend to have greater mechanical advantage. A particularly good indicator of mechanical advantage is the number of load-bearing ropes, which refers to the number of rope strands in the pulley system, not including the pull rope. The number of load-bearing ropes is equal to the mechanical advantage of most of the pulley systems depicted in our problems. Other attributes are less correlated with mechanical advantage. For example, pulley systems with more movable pulleys have a greater mechanical advantage, but knowing the number of movable pulleys without knowing their configuration does not allow one to compute a value for the mechanical advantage. To derive the mechanical advantage of the pulley system from the number of movable pulleys, a person has to know the relevant formula, which in turn depends on the configuration of pulleys in the system. We will refer to attributes that covary with mechanical advantage as relevant attributes.

Figure 1:   A typical pulley problem.

A

B

With which pulley system does the man have to pull with more force to lift the weight?

A
B
If no difference, mark C.

Most of the items in the Bennett test depict two systems that differ on a single attribute: a subject must determine the relationship between that attribute and the effort required to lift the weight. We developed for this study an extended set of pulley problems in which the two depicted systems varied in one or more attributes that were relevant or irrelevant to the system's mechanical function. The problems were designed to examine a number of cognitive processes that were expected to contribute to mechanical ability. For example. problems in which more than one attribute varied between the depicted systems allowed us to measure how subjects combined information from different attributes. Similarly. problems in which both relevant and irrelevant attributes varied allowed us to determine if subjects could distinguish which attributes are relevant to the system's mechanical function. Some of the items in our study. unlike the Bennett items. required quantitative reasoning.

Experiment 1

Method

Problems. We analyze performance on 17 pulley system problems. including some items from the Bennett Mechanical Comprehension Test (Bennett. 1969) and other similar items that were constructed especially for this study. All of the items were multiple choice. requiring a selection among three response alternatives. Each problem depicted two pulley systems lifting a weight and asked which pulley system required more force to lift the weight.

The two depicted pulley systems differed on one or more of the following dimensions: mechanical advantage. weight to be lifted. height (rope length). and pulley size. Pulley systems that differed in mechanical advantage also differed on some subset of the attributes. such as the number of load-bearing ropes and the number of pulleys. that are correlated with mechanical advantage.

Three types of problems were constructed. These three types of problems differed in the kinds of attributes that distinguished the two systems depicted in the problem. In one type of problem. the two depicted systems differed only on attributes irrelevant to the mechanical advantage of a pulley system (height or pulley size). In the second type of problem the two depicted pulley systems differed in mechanical advantage. while the weights they were lifting were equal. In the third type of problem. both the mechanical advantage and the weights were different for the two depicted systems. The three problem types are categorized in Table 1. which lists for each problem type the attributes varied. the number of problems presented. and the ability demonstrated by correct solution of the problems.

Insert Table 1 about here.

In the first set of problems. examples of which are shown in Figure 2. the depicted systems differed only on irrelevant attributes. namely the size of the pulleys or the height of the system. Both the mechanical advantage of the systems and the weights to be lifted were equal. Therefore the effort required to lift the weights in the two cases was equal. These problems allowed us to determine whether a subject could differentiate relevant from irrelevant attributes of pulley systems.

Table 1: Categorization of the Problems in Experiment 1

Attributes Varied	Number of Problems	Ability Demonstrated
Irrelevant Attributes.		
Size of Pulleys	4	Differentiate relevant from irrelevant attributes.
Height	3	Differentiate relevant from irrelevant attributes.
Mechanical Advantage.		
All relevant attributes give correct answer	3	Identify relevant attributes.
Relevant attributes give different answers	3	Prefer attributes more highly correlated with mechanical advantage.
Mechanical Advantage and Weight	4	Compute ratio of weight to mechanical advantage.

Insert Figure 2 about here.

In the second type of problem, the two depicted systems had different mechanical advantage while the weights they were lifting were equal. These problems can be further decomposed into two sub-types. In some of these problems, such as example 3 in Figure 3, the system requiring less effort could be chosen on the basis of some attribute that is correlated with mechanical advantage, such as the number of movable pulleys, the number of ceiling attachments, or the total number of pulleys. These problems allowed us to determine if a subject could choose the correct pulley system on the basis of some relevant attribute. In the other subtype of problems, rules based on different relevant attributes led to different answers to the question. These problems allowed us to identify the attributes of pulley systems that subjects used to make inferences about the relative advantage of different systems, and thus tested whether subjects could differentiate attributes that are highly correlated with mechanical advantage of a system from attributes that are less correlated with mechanical advantage. For instance, in example 4 in Figure 3, a subject who based her answer on the number of pulleys would answer "no difference", while a subject who based her answer on the number of load-bearing ropes would correctly answer that system A requires more effort. Only those subjects who based their judgments on a comparison of the number of load-bearing ropes, or who computed the mechanical advantage by analyzing the balance of forces in the system, would correctly solve all of the problems of this type correctly.

Insert Figure 3 about here.
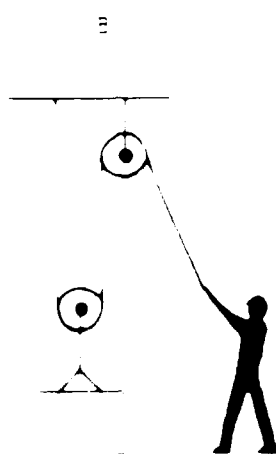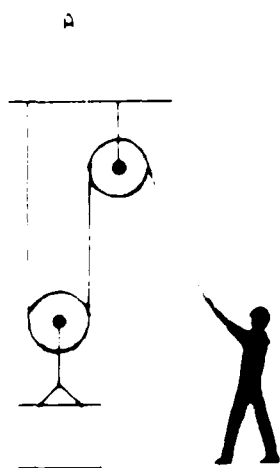

The third set of problems depicted two pulley systems with different mechanical advantage which were being used to lift different weights (see example 5 in Figure 3). These problems tested whether subjects could quantify the mechanical advantage of a pulley system, and whether they knew the correct form of the relation between weight, mechanical advantage and effort. To solve these problems correctly, a subject had to quantify the mechanical advantage of a pulley system and compute the ratio of the weight to the mechanical advantage. The problems in this third set are unlike the items in the Bennett Test, that examine only qualitative knowledge.

Subjects. The subjects were 43 undergraduate students, 27 students at Carnegie Mellon University and 16 students at the Community College of Allegheny County. To ensure a wide range in performance, we included in the sample both students who had taken two or more courses in physics at college level (14) and students who had taken no college level physics courses (29).

Procedure. Thirty-eight subjects were administered the test in a group setting, while five other subjects were tested individually and gave verbal protocols while they solved the problems. Two of the five protocol subjects had taken college level physics.

For the purposes of comparing different levels of ability, the subjects were assigned to one of two groups, a high-scoring group and a low-scoring group, on the basis of their overall scores. A discontinuity in the distribution of scores defined the boundary between the high and low ability subjects. Twenty seven subjects (25 non-protocol subjects and 2 protocol subjects) solved 10 or fewer problems correctly while 15 of the remaining subjects (12 non-protocol subjects and 3 protocol subjects) scored more than 12 of the problems

Figure 2: Examples of problems in which the depicted pulley systems differ on irrelevant dimensions

Example 1:

With which pulley system does the man have to pull with more force to lift the weight?

    A
    B
    If no difference, mark C.

Example 2:

With which pulley system does the man have to pull with more force to lift the weight?

    A
    B
    If no difference, mark C.

Figure 3: Examples of problems in which the depicted systems have different mechanical advantage.

Example 3:

With which pulley system does the man have to pull with more force to lift the weight?

A
B
If no difference. mark C.

Example 4:

With which pulley system does the man have to pull with more force to lift the weight?

A
B
If no difference. mark C.
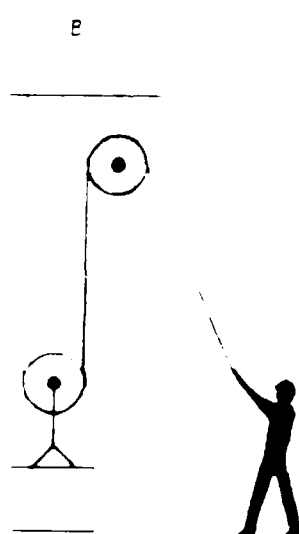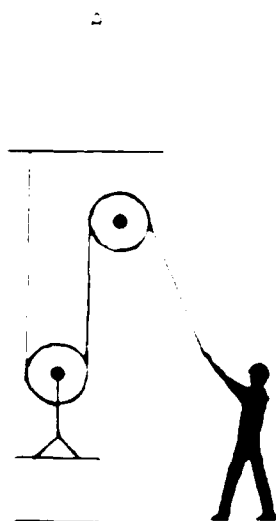
Example 5:

With which pulley system does the man have to pull with more force to lift the weight?

A
B
If no difference. mark C.

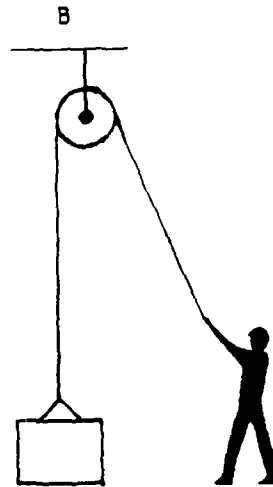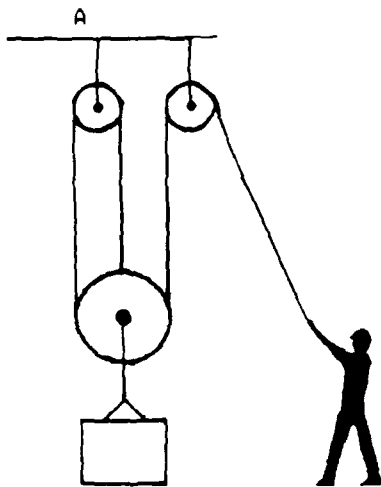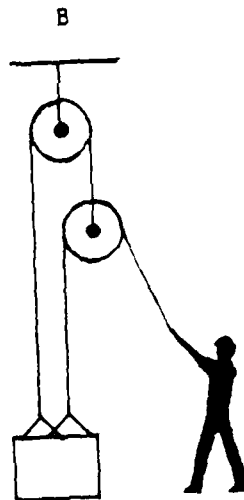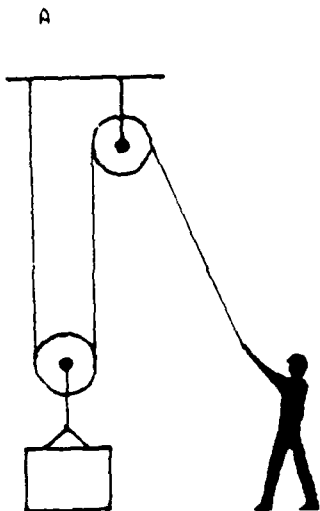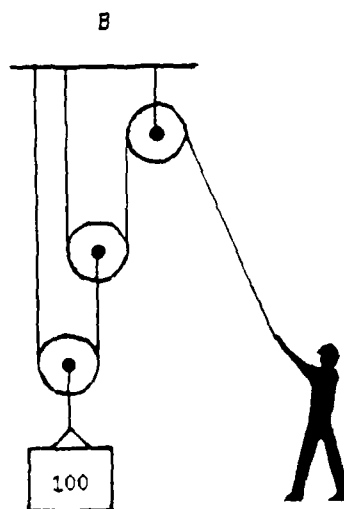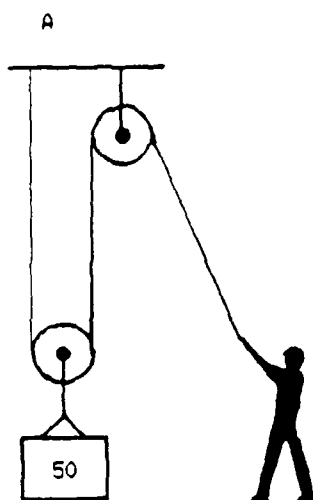correctly. The remaining subject, who solved 11 of the problems correctly, was assigned to the high-scoring group. The high-scoring group therefore consisted of the top third of the distribution. The classification of subjects thus defined was correlated with formal study of physics. Eleven of the thirteen subjects in the high-scoring group had studied college-level physics. Only one of the twenty five subjects in the low-scoring group had studied college-level physics.

Results

We analyzed the data of the 5 subjects who gave verbal protocols separately from the data of the 38 subjects who performed the test in a group setting. The protocols suggested a general account of how subjects solved the test items, which was supported by the group data. The group data also suggested sources of individual differences. We will first consider some solution processes that were general to all subjects, and later focus on the sources of individual differences.

Protocol Analysis. An analysis of the subjects' verbal protocols suggested the following general account of how they solved the test items. The subjects decided which of the two depicted pulley systems' distinguishing attributes (such as the number of pulleys) are relevant to reducing the effort required to lift the weight. They then compared the two systems using rules that relate these attributes to the amount of effort required.

In a typical protocol of a subject solving one of the problems, the subject noted one or more attributes on which the two depicted pulley systems differed, suggested an answer to the problem, and supported his answer with a reason. The attributes that a subject noted were usually attributes of pulley systems that he considered to be causally related to reducing the effort, although subjects occasionally noted a difference in an attribute and stated that it was irrelevant. The reasons that subjects gave for their responses indicated the nature of the relation (which we will call a rule) that they thought existed between an attribute and the effort required. A rule could state that a higher value of some attribute implies a greater effort or that a lower value of the attribute implies a greater effort. For example a subject might think that a system with more pulleys requires a greater effort to lift a weight, or that a system with fewer pulleys requires a greater effort.

We analyzed protocols of 5 subjects solving 17 problems - a total of 85 problem solving episodes. In 73 of these episodes the rule that the subject used could be inferred from the reason that he gave for his answer. In another 6 episodes, the reason given by the subject was ambiguous, and the rule that he used to generate his answer was inferred from the attributes that he noted as relevant, from the rules based on these attributes that he articulated in solving other problems, and from the answer that he gave to the problem. The remaining protocols were unscorable either because the subject's response was too vague or because the subject did not understand the depiction of the pulley system presented in the problem.

The repertoire of rules used by the subjects thus inferred from the five verbal protocols, is presented in Table 2. The rules all pertain to attributes of the visible components of the systems - either their number, size, or attachments to other components. Most of the rules were based on relevant attributes. Three of the protocol subjects thought that the effort required to lift a weight with a pulley system was affected by an attribute that is actually totally irrelevant (size of the pulleys or height of the system).

Insert Table 2 about here.

Some rules were articulated by more subjects than other rules, as shown in Table 2. For example, all five protocol subjects used the rule that a system lifting a lighter weight requires less effort. Four out of the five subjects used the rule that a system with more pulleys requires less effort. The fifth subject considered weight to be the only relevant attribute in all cases. Other rules were more idiosyncratic; for example, two of the five subjects based some answers on the number of rope-to-ceiling attachments in the pulley systems, while the other three subjects ignored this attribute of the pulley systems.

The protocol subjects used a wide variety of rules in solving the problems in which both mechanical advantage and weight were varied. In solving these problems, two subjects used rules based on the ratio between the weight to be lifted and some attribute of the system. Their rules had the correct form, because the effort is the ratio of the weight to the mechanical advantage of the system. However the ratios that these two subjects computed were incorrect because they were based on relevant attributes (the number of pulleys and the number of rope-to-ceiling attachments) that are not perfectly correlated with the mechanical advantage of a pulley system. A third subject thought that system attributes could always compensate for differences in the weight to be lifted, and therefore stated that there was no difference in the effort required for the pairs of pulley systems depicted. The other two subjects based their responses on rules involving either weight or a single attribute of the systems.

In those cases in which one applicable rule dictated one answer and another rule dictated another answer, we were able to infer a subject's preference ordering among his rules from the answer that was ultimately given. Such conflicts arose because many of the rules that the subjects used involve only one attribute, and some of the problems depicted two systems that differ with respect to more than one attribute that subjects considered relevant. In example 4 in Figure 3, a conflict might arise between the rule that a system with fewer ceiling attachments requires more effort, which would produce answer B, and the rule that a system with fewer pulleys requires more effort, which would produce answer C. A subject with a preference for an answer based on the number of ceiling attachments would answer B to this problem.

The preference ordering among rules implies that even if a subject knows a given rule, the rule will not be used to generate the final answer given unless it is the most preferred in a particular situation. For example, even though three of the subjects "knew" the more correct rule that a system with more load-bearing ropes requires less effort, each of these subjects applied this rule on only one problem, because for these subjects, the preference for this rule was weaker than for rules based on other attributes, such as the number of pulleys or the number of rope to ceiling attachments. Another example of a rule preference was some subjects' tendency to prefer rules that indicated a difference between the two depicted systems, as opposed to rules that evaluate the two systems as being the same. For example, the rule that a system with more pulleys requires less effort entails the rule that two systems with the same number of pulleys require the same effort. However the former rule was preferred to the latter, which was rarely mentioned.

Comparison of Problem Types. Comparison of performance on different types of problems of the 38 subjects who solved the problems in a group setting supports the view that subjects apply multiple rules to solving the pulley problems. Given that several different rules are usually applicable in a problem, if subjects are using multiple rules, then

Table 2:   Rules used by the Protocol Subjects in Experiment 1

Rule	Number of Subjects who used the Rule.
A system with ... requires less force:	
less weight	5
more pulleys	4
more load-bearing ropes (tensions)	3
more attachments to the ceiling	3
more free pulleys	2
larger pulleys	2
more fixed pulleys	1
less weight per pulley	1
less weight per attachment	1
less height	1
mixed pulley sizes	1

performance should be better in those problems in which the different applicable rules converge on the correct answer. This prediction was confirmed. The three problems in which the applicable rules converged on the correct answer had a significantly higher proportion of correct responses than the three problems in which some of the rules conflicted (.51). as indicated by a t-test for matched pairs ($\underline{t}$(37) = 2.11, $\underline{p}$ < .05). The lower proportion of correct responses in problems where rules conflicted indicated that some erroneous responses were generated when subjects preferred a rule based on a relevant attribute that is imperfectly correlated with mechanical advantage.

Comparison of performance on different types of problems also suggests that subjects had particular difficulty with problems that required the quantitative combination of attributes. The four problems that required quantitative understanding of pulley systems (i.e.. ratios) produced a lower proportion of correct responses (.43) than the six problems requiring only qualitative knowledge (.58). $\underline{t}$(37) = 3.04. $\underline{p}$ <.001).

In summary, subjects answered the test items by choosing the response alternative dictated by the most preferred rule that was applicable in that item. The rules pertained to visible attributes of the pulley systems and could be either irrelevant or relevant to the systems' mechanical function. Rules based on attributes that were less correlated with mechanical advantage were often preferred to rules based on attributes that were more correlated. Problems in which different rules led to different answers produced more errors than problems in which different rules converged on the correct answer.

Individual Differences. The response patterns of most of the 38 subjects who solved the problems in a group setting could be classified as consistent with the rules observed in the protocols. That is. most of the 38 subjects chose answers as though they were using some subset of the rules manifested in the protocols. These response patterns provided information on the sources of individual differences in mechanical ability.

Three abilities accounted for individual differences in different subsets of problems in the test. These were (1) ability to discriminate relevant from irrelevant attributes. (2) consistency of rule use and (3) ability to quantitatively combine information about two attributes within a single rule. We will discuss each of these abilities in turn.

High-scoring subjects were better able to discriminate relevant from irrelevant attributes of pulley systems. In problems in which the irrelevant attributes of height and pulley size were varied. the majority of high-scoring subjects correctly identified these attributes as irrelevant (see Table 3). This was reflected in the answers that they chose. High-scoring subjects chose a significantly higher proportion of correct responses (.90) than did low-scoring subjects (.44) in the three problems that varied the height of the system. as indicated by a two-sample t-test ($\underline{t}$(33) = 4.51, $\underline{p}$ <.001). Similar results were found in the case of the four problems that varied pulley size, where .98 of high-scoring subjects' responses and .35 of low-scoring subjects' responses were correct ($\underline{t}$(27) = 8.29, $\underline{p}$ <.001). The majority of those subjects who considered height or pulley size to be relevant were consistent in the rule that they used. so it was possible to classify the responses of almost all subjects to these problems as rule governed (see Table 3).

Insert Table 3 about here.

The responses of high-scoring subjects on problems that varied mechanical advantage were both more likely to be consistent with one of the rules identified in the protocols and more likely to be correct. If consistency is defined as having at least five out of six

Table 3: Classification of Responses to Problems in which the depicted Pulley Systems differ on Irrelevant Attributes.

Rule	Number of Subjects who used the Rule.	
	High-scoring	Low-scoring
(1) Height of system varied.		
Height is irrelevant	11 (85%)	10 (40%)
A system with ... requires less effort		
less height	0 (0%)	7 (28%)
more height	1 (8%)	6 (24%)
Total Classified:	12 (92%)	23 (92%)
(2) Size of pulleys varied.		
Pulley size is irrelevant	13(100%)	7 (28%)
A system with ... requires less effort		
larger pulleys	0 (0%)	8 (32%)
smaller pulleys	0 (0%)	3 (12%)
Total Classified:	13(100%)	18 (72%)

responses that are consistent with some rule. twelve of the thirteen high-scoring subjects responded consistently with one of the rules (see Table 4). In contrast low-scoring subjects were less consistent in their use of rules. Even if consistency is defined more leniently. as having four of their six responses consistent with a rule. only eleven of the twenty-five low-scoring subjects could be called consistent. Thus. low-scoring subjects are either less consistent in their application of rules to a given problem or less consistent in their resolution of conflicts between different rules. The answers of high-scoring subjects were not only more consistent with each other. but were also more consistent with correct rules. Consequently. high-scoring subjects answered a significantly higher proportion (.77) of these *six problems correctly* than did low-scoring subjects (.47). (t(36) = 4.48. p<.001). Consistency of rule use is difficult to assess in the absence of verbal protocols. so one of the purposes of Experiment 2 was to provide protocols from a larger number of subjects.

Insert Table 4 about here.

High-scoring subjects were also more likely to demonstrate quantitative understanding of pulley systems than were low-scoring subjects. as shown in Table 4. In problems involving both mechanical advantage and weight differences. the responses of ten of the twelve high-scoring subjects were consistent with rules expressing a ratio of the weight to some attribute of the system. such as weight per load-bearing strand. attachment. or pulley. As observed in the protocols. it is likely that these ratios were sometimes based on incorrect indices of mechanical advantage. such as the number of ceiling attachments or pulleys. The low-scoring subjects. on the other hand. were more likely to base their comparisons of the systems either on weight or on a single attribute of the system. but typically did not combine the consideration of weight and the system attribute into a single rule. The most common rule used by these subjects was that more effort is required to lift a heavier weight. High-scoring subjects answered a *much higher proportion (.62) of these four problems correctly* than did low-scoring subjects (.33) (t(36) = 4.03. p<.001).

Relation of Specific Abilities to Total Performance. The response patterns indicated that high-scoring subjects are better able to identify the attributes relevant to the operation of a pulley system. that they are more consistent in their use of rules. and that they are more likely to use rules that indicate a quantitative understanding of pulley systems. Not only do these three abilities have significant effects on performance. but they are also similarly related to the total scores. as assessed by the following procedure. Each subject was given a score of 1 or 0 corresponding to each of the three abilities. A score of 1. based on the response pattern on the relevant problems. indicated that the subject had the ability in question. while a score of 0 indicated that the subject did not have this ability. Each ability score had a correlation with the overall score that lay between .49 and .51. Thus the three abilities are of approximately comparable importance in predicting an individual's performance. Together the three abilities accounted for 38.6% of the variance among the total scores.

Discussion

The results of Experiment 1 suggested a characterization of the processes used to solve the items in a test of mechanical ability. According to this characterization. individuals decide which attributes of pulley systems are relevant to reducing the effort required to lift a weight: they compare the depicted pulley systems by applying rules that relate these attributes to the effort that must be exerted. When several different rules are applicable in a given situation. preferences among these rules determine which rule is used

Table 4: Classification of Responses to Problems in which the Pulley Systems have different Mechanical Advantage.

Rule	Number of Subjects who used the Rule.	
	High-scoring	Low-scoring

(1) Systems which differ on M.A.

A system with ... requires less force:

more load-bearing ropes,

or less weight per load-bearing strand	7 (54%)	7 (28%)
more pulleys, or less weight per pulley	3 (23%)	3 (12%)

more attachments, or less weight

per attachment, or more movable pulleys	2 (15%)	1 (4%)
Total Classified:	12 (92%)	11 (44%)

(2) Systems which differ on M.A. and Weight.

A system with ... requires less force:

less weight per load-bearing rope	5 (38%)	0 (0%)
less weight per attachment	3 (23%)	3 (12%)
less weight per pulley	2 (15%)	2 (8%)
less weight	1 (8%)	8 (32%)
more ropes	1 (8%)	2 (8%)
more attachments, or more movable pulleys	0 (0%)	2 (8%)
Total Classified:	12 (92%)	17 (68%)

to generate an answer to the problem.

The experiment also provided information on the sources of differences between individuals in performance on tests of mechanical ability. The response patterns indicated that high-scoring subjects are better able to correctly identify the attributes relevant to the operation of a pulley system. that they are more consistent in their use of rules. and that they are more likely to use rules that indicate a quantitative understanding of pulley systems.

A Model of Performance.

In order to specify mechanisms that can underlie performance on the problems and that can account for the individual differences identified in Experiment 1. we developed a simulation model. The model simulates the performance of one high-scoring and one low-scoring protocol subject from in Experiment 1. It simulates the response choices that the subjects gave to the problems in Experiment 1. as well as stating the rationale for each choice.

The simulation model is written in the Soar production system language (Laird. Newell. and Rosenbloom. in press). As in other production systems. Soar's procedural knowledge is contained in productions. some of which. in this case. are intended to correspond to the rules subjects use in solving the pulley problems. One property of Soar that makes it particularly suitable for modeling performance in the pulley problems is its built-in ability to manipulate goals. Soar's processing is driven by a top-level goal in a problem space. as well as by sub-goals that Soar itself formulates as necessary to fulfill a higher level goal. The top-level goal in the pulley problems is to find out which one of the two depicted pulley systems requires more effort to lift its weight. Soar uses several heuristic methods to generate a subgoal when a current goal cannot be satisfied directly. Another property of Soar. useful in this domain. is that when more than one of several rules (productions) are applicable in a given situation. Soar can choose among them on the basis of a preference ordering. In our model. this preference ordering is intended to correspond to the rule preferences exhibited by the human subjects.

Problem Representation. The model operates on a problem description for each of the 17 problems in Experiment 1. Each problem description contains all the information that is directly available to a human subject through visual inspection. However. not all of the information in the problem description is necessarily used by the model or by the subject it simulates.

The format of a problem description is a structured description list that consists of identifiers and lists of attributes and values. There are four types of attributes: properties. relations. comparisons. and questions. The simplest type of attribute is a property of a pulley system or component of a pulley system. such as the number of pulleys in a system. or the height of a system. The second type of attribute is a relation between two objects. A relation names a source object and a related object. states the type of relation it is. and contains a value for the relation. For example. a relation might state that a particular pulley is fixed to the ceiling. The third type of attribute. a comparison. compares two properties or two relations. For example. a comparison might state that the height of pulley system A is greater than the height of pulley system B. The fourth type of attribute. a question. contains an attribute with a missing value and states that the value should be obtained. The requirement in each item of the test. namely to compare the relative efforts required to lift the weights with the two depicted pulley systems. is represented as a question about the comparison of the effort attribute.

Production Rules. The simulation model uses a set of productions that can be divided into two subsets. one subset common to all subjects. and a second subset unique to the individual whose solutions were simulated. The common productions constitute a conventional production system model of problem-solving. with many of the conventional problem-solving mechanisms provided by Soar itself. The common productions control the operators that seek information about the problem and the operators that generate answers to the question posed. express the reasons for producing these answers. and stop the processing when the final answer has been selected.

The subject-specific productions determine what information a particular individual seeks and how he reasons from that information to generate an answer to the problem. These productions reflect the rules that the subject possesses relating attributes of pulley systems to their function (reducing the effort required to lift a weight). The conditions of these productions specify the situations in which each rule can be applied. The conditions of application of a rule can include values of properties. relations. or comparisons. For example. the production that determines that information about the number of pulleys in the two systems should be sought might be evoked when the values for the two effort attributes are missing and the weight is the same for the two pulley systems.

The model can evoke one of two types of operators. elaboration operators and hypothesis operators. When a value in a question is missing. elaboration operators look for information in the problem statement that might be relevant to answering the question. For example. if the question asks for a comparison of the effort attributes of two systems. an elaboration operator might look for the values of the weights to be lifted by the two systems. Hypothesis operators suggest values for attributes that are sought by elaboration operators and use these values to suggest tentative answers to the problem (we will use the term "suggestion" to refer to a tentative answer). A suggested answer can state that a higher value of the attribute implies a greater effort or that a lower value of the attribute implies a greater effort. Each suggested answer is accompanied with a reason for this answer. For example. a hypothesis operator may suggest pulley system A requires a greater effort than system B because the weight that system A is lifting is heavier. The alternation between operators that elaborate the current knowledge state and those that act in light of the elaboration is a part of the Soar architecture.

The productions choose among elaboration and hypothesis operators on the basis of preferences. expressed in Soar as special data elements. A preference might favor an answer supported by a particular reason. For example a preference might favor an answer based on the amount of weight to be lifted by a system over an answer based on the number of pulleys in a system. Alternatively. a preference might express a bias for a particular response. For example. a preference might favor a hypothesis operator stating that the efforts required to lift the loads of the two pulley systems are different over an operator stating that the efforts are the same.

The model proceeds from the problem description and question to its ultimate response by evoking a sequence of operators that derive information from the problem description and suggest answers on the basis of the obtained information (see Figure 1). When the question is first interpreted. an elaboration operator is evoked to seek the information that the question interrogates. The question ("With which pulley system does the man have to pull with more force to lift the weight?") initiates a comparison of the effort attributes of the two pulley systems. Because there is no information available that allows this comparison to be made directly. additional elaboration operators are evoked to seek additional information that might be relevant to the answer. For example. information about the number of pulleys or ceiling attachments in the two pulley systems might be

sought at this point. In addition. if the person being modeled has sufficient knowledge to calculate the efforts required by the two pulley systems. a subgoal to calculate the efforts is generated. (The dotted lines in Figure 4 indicate components of the model that are present for subjects with this knowledge). Hypothesis operators use the information obtained by elaboration operators to suggest answers to the question. If no answer is suggested. the model chooses randomly among the possible answers. If only one answer is suggested. it becomes the response of the model for that problem. If more than one answer is suggested. a subgoal is created to resolve the tie. To satisfy the subgoal of resolving the tie. one hypothesis operator may be selected over another as a result of a preference. Otherwise. a random choice is made among the operators.

A worked out example helps to illustrate how the model operates. Say that the model was simulating performance on problem example 4 in Figure 3 by a subject with the rules "a system with more pulleys requires less force" and "a system with more attachments to the ceiling requires less force". The model would first evoke an elaboration operator to compare the efforts required to lift the weight with the two pulley systems. When this information was not found directly in the problem. a subgoal would be created to find information relevant to comparing the efforts. and at this stage elaboration operators would be successful in finding comparisons of the number of pulleys and the number of ceiling attachments in the problem statement. Hypothesis operators would then produce two suggested answers to the problem. one stating the two pulley systems require the same effort because they have the same number of pulleys. and another stating that pulley system A requires less effort because it has more pulleys. At this stage a preference. say for the answer based on the number of pulleys. might resolve the conflict between the two suggested answers. The elaboration operator to compare the efforts of the two systems would be successful in finding a comparison (i.e. that the two pulley systems require the same effort) and this would become the answer of the model for that problem.

Insert Figure 4 about here.

Modeling individual differences in performance.

The model simulates the performance of individual subjects on the 17 pulley problems included in Experiment 1. In this section. the simulation of one high-scoring subject and one low-scoring subject will be contrasted.

The three sources of individual differences observed in Experiment 1 are modeled in the simulation in the following ways. To account for the differences among subjects in what they consider to be relevant. the model for a given subject relates the effort required in the case of a particular pulley system to precisely those attributes of the system that the subject considers relevant. That is. the attributes that were considered relevant were in the conditions of the productions embodying the mechanical rules. To account for the differences among subjects in how consistently they use rules. the model varies or keeps constant its preferences among hypothesis operators across the different problems. If there is a preference for one hypothesis operator over all other hypothesis operators in a situation. the model will always choose the answer and the reason given by that operator in any similar situation. If there is no preference among operators. then the model chooses randomly among applicable operators. producing the same type of inconsistent behavior as observed for low-scoring subjects in Experiment 1. Finally. to account for the differences among subjects in their ability to quantitatively combine information from two relevant attributes. the model can either contain or not contain productions that suggest values for the effort based on a ratio of the weight of the system to some other relevant attribute.

Figure 4: Flow of control of the simulation model through a problem.


```
                    ┌─────────────────┐
                    │  Read Question  │
                    └─────────────────┘
                       │            ┊
                       ▼            ▼
         ┌──────────────────┐   ┌··················┐
         │  Set subgoal to  │   :  Set subgoal to  :
         │  compare efforts │   :  calculate efforts:
         └──────────────────┘   └··················┘
                   │                    ┊
                   └──────────┬─────────┘
                              ▼
                  ┌──────────────────────┐
            ┌────▶│  Seek value of pulley│
            │     │   system attribute   │
            │     └──────────────────────┘
            │                 │
            │                 ▼
            │     ┌──────────────────────┐
            │     │ Suggest value for    │
            │     │ attribute from       │
            │     │ problem description  │
            │     └──────────────────────┘
            │                 │
            │                 ▼
            │             ╱ Is  ╲          no    ┌──────────────────┐
            │            ╱attribute╲──────────────▶│ State attribute  │
            │            ╲relevant?╱               │  is irrelevant   │
            │             ╲      ╱                 └──────────────────┘
            │               yes                            │
            │                 ▼                            │
            │             ╱  Any  ╲                        │
            │     yes    ╱distinguishing╲                  │
            └───────────╱  attributes  ╲◀────────────────┘
                        ╲    left?     ╱
                         ╲            ╱
                              no ┊······················┐
                               │                        ▼
                               │              ┌··················┐
                               │              :  Calculate efforts:
                               │              └··················┘
                               ▼                     ┊
                  ┌──────────────────────┐           ┊
                  │  Suggest values for  │◀···········┘
                  │  effort comparison   │
                  └──────────────────────┘
                              │
                              ▼
                  ┌──────────────────────┐
                  │   Select one value   │
                  └──────────────────────┘
                              │
                              ▼
                  ┌──────────────────────┐
                  │     State answer     │
                  └──────────────────────┘
```

Model of a High-scoring Subject. The simulation model was intended not only to make the same pattern of responses as the human subjects, but also to base its answers on the same reasons, and to state those reasons. As a result, in choosing which subjects to simulate, we chose from among the five protocol subjects. The high-scoring subject that was chosen provided particularly clear explanations of his answers. Unfortunately his score, 12 correct out of 17, was not as high as we would have liked for a "high-scoring" subject, and was only marginally better than the performance of the low-scoring subject that we simulated (who solved 10 of the 17 problems correctly). However, two of his errors seemed attributable to encoding errors, in which he incorrectly counted the number of pulleys in a system. Without these superficial errors, the subject would probably have answered 14 out of the 17 problems correctly. We simulated the incorrect encoding by giving the simulation for this subject the same incorrect values for the number of pulleys in these problems as were encoded by the subject.

The simulation was given productions expressing the rules and preferences that we inferred from the high-scoring subject's protocol and in 16 out of the 17 problems, the model provided both the same response and the same explanation of the response as the subject. For example, in solving the pulley problem shown in Figure 1, which asks which of two pulley systems requires more force to lift the same weight, the high-scoring subject gave the following explanation for his answer:

> It would be the second man because he has one less pulley helping him,
> and the more pulleys you have, the easier it is to lift something. So
> I'll say B."

The simulation also answered B for this problem and gave the reason that pulley system A has more pulleys than B. In the single instance in which the simulation responded differently from the subject, the subject made a one time application of a rule that was not accommodated by the simulation. The high-scoring subject was otherwise very consistent in the application of a small set of rules.

Table 5 summarizes the rules required to simulate the high-scoring subject's solutions. The first set of rules (1a, b, and c) is concerned with elaboration operators that attempt to obtain information about the effort or about attributes relevant to the effort. The operator for finding out about the efforts directly (1a) is included in the model because the subject being simulated did occasionally calculate the efforts (albeit incorrectly). The operators for finding out about the number of pulleys, load-bearing ropes, and weights (1b), generate the information that provides the basis for reasoning about the relative efforts. Also included are operators for finding out about irrelevant attributes, namely the sizes of pulleys, distances between pulleys, and heights of the pulley systems (1c). These elaboration operators are included because the subject noticed differences in these attributes but did not relate the differences to the effort. Similarly, the model notices these differences but subsequently no hypothesis operators make suggestions on the basis of these attributes.

Insert Table 5 about here.

Rules that make suggestions (rules 3a, b, and c in Table 5) generate a tentative answer (hypothesis operator) to each problem on the basis of comparisons of the number of pulleys, the weights, or the number of load-bearing ropes. The answers that the subject gave suggest that he considers the effort required to lift a weight with a particular system to decrease with the number of pulleys and load-bearing ropes, and increase with the

**Table 5 : Summary of rules required to simulate the
High-Scoring Subject's Solutions.**

1. Rules for evoking elaboration operators to seek information or notice differences
   (9 productions).

   a. Evoke operators that calculate the effort required to lift the load with each
      pulley system.

   b. Evoke comparisons of attributes that permit suggestions about the effort
      comparison.

      i. Comparison of number of pulleys.

      ii. Comparison of number of load-bearing ropes.

      iii. Comparison of weights.

   c. Evoke operators for differences that are noticed but are ultimately
      considered irrelevant.

      i. Comparison of the size of corresponding pulleys.

      ii. Comparison of the distances between corresponding pairs of pulleys.

      iii. Comparison of the distances from the effort to the first pulley (i.e.
           the height of the pulley system).

2. Rules for sequencing elaboration operators and terminating the search for some
   attributes (4 productions)

   a. Seek comparisons between pulleys and weights before calculating efforts. if
      either pulleys or weights are equal. base the comparison on an unequal
      attribute without calculating efforts.

   b. If the efforts are calculated. terminate the search for information about all
      other attributes.

3. Rules to make suggestions (i.e. hypothesis operators) about the effort comparison
   (5 productions).

   a. The effort decreases with the number of load-bearing ropes.

   b. The effort decreases with the number of pulleys.

   c. The effort increases with the weight.

Table 5 : Summary of rules required to simulate the
High-Scoring Subject's Solutions (continued).

4. Rules for combining suggestions (1 production).

    a. The effort of each system is calculated by dividing the weight by the
number of pulleys in the system.

5. Rules for choosing among suggestions. (i.e. selecting tied hypothesis operators) (3
productions).

    a. Prefer suggestions indicating a difference to those predicting equality.

    b. Prefer weights over pulleys as a reason for an answer of no difference.

weight, and the hypothesis operators in the model function likewise.

The high-scoring subject has a simple set of preferences that are used in choosing among competing hypothesis operators or among the answers they produce. For example, he prefers an answer of a difference in the effort required for the two pulley systems rather than an answer of equality. If all the operators indicate equality, he prefers an answer based on the weights to one based on the number of pulleys. These types of preferences, observed in the subject's responses and protocol, are directly represented as such in the model (rules 5a and b in Table 5).

The model of the high-scoring subject contains productions that control the sequence of search for information (rules 2a and b in Table 5). For example, the model (and the human subject) seeks information about the weight and the number of pulleys before evoking the operator that attempts to calculate the efforts directly. This is adaptive for two reasons. First, if only the weights or only the numbers of pulleys are equal, a value for the effort comparison can be suggested in the basis of whichever of these attributes is not equal, and there is no need to calculate the effort. Second, if the effort does have to be calculated, the information obtained from the earlier comparisons of weights or pulleys can be used in calculating the effort, by dividing the weight by the number of pulleys (Rule 4a in Table 5). Once the effort values are calculated, all other elaboration operators are terminated because a direct calculation is assumed to provide the answer. Thus, the high-scoring subject's knowledge is organized in a way that provides an efficient search for information.

Model of a Low-scoring Subject. Of the three low-scoring subjects who gave protocols, we chose to simulate the one whose protocols were clearest about the alternative answers and the supporting justifications that were being considered in each problem. This subject was also relatively consistent in applying rules to solve the problems. She answered 10 of the 17 problems correctly.

The simulation model was given productions expressing the rules and preferences that we inferred from the low-scoring subject's protocol, and was able to match the answers of this subject in 16 of the 17 problems. In addition, it matched the reasons for her answers on these problems. In 3 of the 16 matching cases, either the simulation or the subject gave an additional answer for the question, not given by the other. For example, in solving the pulley problem in Figure 1, the subject gave the following explanation for her answer.

"I would say B has to pull more because there's only one
attachment [of a rope] the ceiling, so that makes it harder."

The simulation also gave the answer B to this problem, and gave as reasons that pulley system A has both more rope-to-ceiling attachments and more pulleys than system B. The explanation given by the model is reasonable, because in two other problems of this type, the subject gave both of these reasons for her answer. In the single problem in which the simulation did not match the subject's answer, the subject gave a vague reason for her answer from which it was not possible to determine what rule she was using.

The rules of the low-scoring subject's model are more complicated in some respects but simpler in other respects than the rules of the high-scoring subject's model. Table 6 lists some of the key productions in the model of the low-scoring subject. There are seven productions that evoke elaboration operators, two fewer than in the high-scoring subject's simulation. The two productions that are absent in the low-scoring model are the ones that try to find out directly about the effort values. Comparisons of several attributes are

evoked: some are relevant attributes (weights, number of pulleys, number of rope-to-ceiling attachments) while others are irrelevant attributes (pulley size, distance between pulleys). One of these attributes was not noticed in the high-scoring model, namely rope-to-ceiling attachments.

Insert Table 6 about here.

Just as the high-scoring subject uses the number of pulleys to index the mechanical advantage of a pulley system, the low-scoring subject uses rope-to-ceiling attachments as an indicator of a pulley system's advantage, although she does not make a numerical division to calculate effort values. That is, the low-scoring subject believes that rope-to-ceiling attachments make the ceiling bear some of the load. This rule somewhat resembles determining the number of load-bearing ropes, and sometimes produces the correct answer.

The low-scoring subject does not organize the search for information (there are no rules corresponding to rules 2a and b in Table 5 describing the high-scoring simulation); in the model, the sequence of elaboration operators is randomly determined by Soar.

The low-scoring simulation produces more suggestions than the high-scoring simulation. This is because the low-scoring model makes suggestions about an irrelevant attribute in addition to making suggestions about a number of relevant attributes. Within each pulley system, the effort is hypothesized to be related to three relevant attributes (number of pulleys, number of rope-to-ceiling attachments, and weight) and to one irrelevant attribute (pulley size).

Because the low-scoring subject does not know the correct relations between the attributes of a pulley system and the weight and the effort, she must find some way to produce a response by combining the suggestions generated by several individual rules. In general, if two or more suggested answers are the same and the explanations associated with them are different but mutually compatible, the human subject combines them into a single answer justified by the several explanations. The simulation for this subject models this aspect of her performance. It combines suggestions both when they are based on similar explanations, such as the sizes of two different pairs of corresponding pulleys, and when they are based on less similar explanations, such as numbers of pulleys and the sizes of weights. If two or more suggestions produce contradictory answers, these comparisons cancel each other and the low-scoring simulation gives an answer of equality.

Like the high-scoring simulation, the low-scoring simulation prefers an answer of a difference between the pulley systems over an answer of equality. Suggestions based on several attributes are preferred to those based on a single attribute.

Discussion

The model contributes to the understanding of mechanical ability by specifying both the shared mechanisms that underlie performance of all subjects, and the mechanisms that account for differences in performance between subjects. It successfully simulates the performance of a high-scoring and a low-scoring subject, demonstrating its ability to account for the difference in performance among individuals.

There were some similarities between the high-scoring and the low-scoring subject. The same general model accounted for the performance of the two subjects. This model interpreted the question posed in a problem as requesting a comparison of the effort attributes of the two pulley systems depicted. It then sought information from the problem description that might lead to answers to the question. Finally, it resolved conflicts between

Table 6: Summary of Rules required to simulate the
Low-Scoring Subject's Solutions.

1. **Rules for evoking elaboration operators to seek information or notice differences (7 productions).**

    a. Evoke comparisons of attributes that permit suggestions about the effort comparison.

        i. Comparison of number of pulleys.

        ii. Comparison of number of rope to ceiling attachments.

        iii. Comparison of weights.

        iv. Comparison of the size of corresponding pulleys.

    b. Evoke operators for differences that are noticed but are ultimately considered irrelevant.

        i. Comparison of the distances between corresponding pairs of pulleys.

2. **Rules for sequencing elaboration operators and terminating the search for some attributes (0 productions).**

3. **Rules to make suggestions (i.e. hypothesis operators) about the effort comparison (9 productions).**

    a. The effort decreases with the number of pulleys.

    b. The effort decreases with the number of rope-to-ceiling attachments.

    c. The effort increases with the weight.

    d. The effort decreases withd the size of corresponding pulleys.

    e. A system with mixed pulley sizes requires a greater effort than one with equal pulley sizes.

Table 6: Summary of Rules required to simulate the
Low-Scoring Subject's Solutions (continued).

4. Rules for combining suggestions (9 productions).

   a. If there are multiple pulleys in a system. and the suggestions about the
       effort comparison based on pulley sizes have the same value. combine
       them into a single suggestion.

   b. If multiple suggestions have opposite predictions for the effort comparison.
       they cancel each other and an equal suggestion is created.

   c. Combine suggestions based on pulleys. rope-to-ceiling attachments. and
       weights if their predictions are the same. and they predict a difference.

   d. Combine suggestions based on pulleys and weights if their predictions are
       the same and they predict no difference.

5. Rules for choosing among suggestions (i.e. selecting tied hypothesis operators) (2
   productions).

   a. Prefer suggestions indicating a difference to those predicting equality.

   b. Prefer combined suggestions.

suggested answers and produced an answer to the problem. Both of the subjects whose performance was simulated based their answers on visible attributes of pulley systems. In addition, both subjects had a preference for answers indicating a difference between the systems depicted over answers indicating equality.

The simulations for the high-scoring and the low-scoring subjects also differed in a number of respects. All of the answers suggested by the simulation of the high-scoring subject were based on relevant attributes while some of the answers suggested by the low-scoring simulation were based on an irrelevant attribute. The high-scoring simulation sought to determine the efforts required to lift the loads of the pulley systems directly and had the ability to calculate values for the efforts by quantitatively combining two attributes: weight and number of pulleys. The low-scoring simulation did not attempt to determine the efforts directly and did not quantitatively combine attributes. Finally, the high-scoring simulation organized the search for information so that the efforts were calculated directly only when the answer could not be determined by means of a simpler comparison. The order of search for information in the low-scoring model was random.

The models suggest what types of mechanisms might be underlying the three sources of individual differences identified in Experiment 1. Differences in what subjects consider relevant are accounted for by differences in what information is used to make suggestions about the effort comparison. Differences in consistency among individuals are accounted for in terms of the presence or absence of preferences. If there is a preference for one hypothesis operator over another, the model's responses will be consistent over problems. If there is no preference, the choice among operators will be random and produce inconsistent behavior. Quantitative knowledge is accounted for by the existence of productions that suggest answers based on a ratio of the weight to some other relevant attribute. This quantitative knowledge allows values for the effort attribute to be calculated directly.

Although the model is designed to simulate steady-state performance, it suggests how mechanical ability may develop. Two mechanisms of the simulation model can explain how a rule can gradually come to be used correctly. First, a correct rule may initially have excessively restrictive conditions of application: with repeated use it can be gradually generalized to its full and correct range by omitting the overly restrictive conditions. A second mechanism to account for the gradual emergence of a correct rule is a change in the preferences among several rules. A barely-acquired correct rule may start with an initially low preference, but each time it succeeds in generating the correct answer, its preference index, and hence its frequency of use, may increase. These procedures for gradually acquiring a correct rule are not implemented in the model at present.

Experiment 2.

The results of the first experiment suggested a general account of the subjects' performance and some sources of individual differences. Experiment 2 was run in order to further elaborate some of the findings of Experiment 1. Experiment 2 included problem types that varied new combinations of attributes. In addition, all subjects in Experiment 2 were required to give verbal protocols, so that their solution processes could be analyzed more directly than was possible from response pattern data.

Experiment 2 examined the resolution of conflicts between a rule involving a relevant attribute and a rule involving an irrelevant attribute. In Experiment 1, the two systems always differed either with respect to a relevant attribute or with respect to an irrelevant attribute, but not both. It is possible that people of low ability consider an irrelevant attribute relevant only if it is the only distinguishing attribute. However, in the presence of a relevant distinguishing attribute they might prefer a rule based on the relevant attribute. An alternative and equally plausible hypothesis is that low ability subjects mistakenly think

that the irrelevant attributes are relevant regardless of any other variation. If the latter hypothesis is correct. then subjects of low ability should have no consistent preference for rules based on relevant attributes over rules based on irrelevant attributes. To test these hypotheses. Experiment 2 included some problems in which both relevant and irrelevant attributes varied between the two systems.

Experiment 2 also introduced some new problems that tested how subjects took some configurational properties of a pulley system into consideration. The rules identified in Experiment 1 were based on attributes of system components. such as the number of pulleys or the length of the ropes. However the mechanical advantage of a pulley system depends. not just on the components it contains. but also on how these components are configured. The new problems tested whether subjects take configuration into account by examining their treatment of an extra pulley. which was irrelevant because of the way the system was configured.

In Experiment 2. all subjects were asked to give verbal protocols while solving the problems. Protocols can elaborate on the information available from response patterns in three ways. First. they can indicate which rule(s) a subject is using when a number of different rules can produce the same pattern of responses. Second. protocols can indicate how subjects resolve conflicts when two or more rules suggest conflicting answers. Third. they can indicate whether the subject has some type of understanding of the problems that does not produce a consistent pattern of responding. The protocols proved to be particularly informative in this regard in problems varying both weight and mechanical advantage. problems that often produced inconsistent response patterns.

Method

Problems. Performance on 44 pulley problems was analyzed. The problems had the same format as those in Experiment 1. The attributes on which systems differed were height. mechanical advantage. and weight to be lifted.

The major classification of problems was similar to Experiment 1. There were three basic types of problems. In the first type of problem. the two depicted pulley systems differed on an irrelevant attribute. In the second type of problem the two pulley systems differed on attributes that can affect mechanical advantage (relevant attributes). In the third problem the two pulleys systems differed in mechanical advantage and weight to be lifted. The problem types used are shown in Table 7. which lists for each problem type the attributes varied. the number of problems. and the knowledge indicated by correct solution of the problems.

Insert Table 7 about here.

The first type of problem depicted two pulley systems that differed only on the irrelevant attribute of height. or. on height as well as on a relevant attribute. These problems allowed us to determine subjects' preferences among rules based on irrelevant attributes of pulley systems versus rules based on relevant attributes. Four problems varied only the height of the pulley system while eight problems varied both the height and the mechanical advantage of the pulley system. and the remaining eight problems varied both the height of the pulley system and the weight to be lifted. Problems in which two attributes were varied were designed so that half of the problems would involve conflict for a subject who thought that a system with greater height requires more effort. while the

Table 7:   Categorization of the Problems in Experiment 2

Attribute(s) Varied	Number of Problems	Ability Demonstrated
Irrelevant Attributes.		
Height	4	Differentiate relevant from irrelevant attributes.
Height and M.A.	8	Prefer relevant to irrelevant attribute.
Height and Weight	8	Prefer relevant to irrelevant attribute.
Attributes relevant to Mechanical Advantage.		
All relevant attributes give correct answer	4	Identify relevant attributes.
Relevant attributes give different answers	4	Prefer attributes highly correlated with mechanical advantage.
Irrelevant pulley included	4	Identify irrelevant pulley from the system Configuration.
Mechanical Advantage and Weight.		
M.A. compensates for weight	4	Compute ratio of weight to M. A.
M. A. difference greater	4	Compute ratio of weight to M. A. or prefer M.A. to weight.
Weight difference greater	4	Compute ratio of weight to M. A. or prefer weight to M.A.

other half would involve conflict for a subject who thought that less height requires more effort.

The second type of problem depicted two systems that differed in relevant attributes. This second type can be decomposed into three sub-types. The first sub-type were problems in which a number of rules based on relevant attributes, such as the number of pulleys or the number of ceiling attachments, converged on the correct answer. These problems allowed us to determine if a subject could compare pulley systems on the basis of some relevant attribute. The second sub-type were problems in which two or more rules, based on different attributes, led to different answers. Performance on these problems revealed subjects' preferences among rules based on different relevant attributes, some of which were more highly correlated with mechanical advantage than others. The third sub-type of problems depicted pulley systems that differed in an attribute (number of pulleys) that is usually correlated with mechanical advantage, but in this sub-type of problem the two depicted systems did not differ in mechanical advantage. In these problems, one of the systems included an extra pulley, attached either to the ceiling or floor, that did not affect mechanical advantage (see Figure 5). A subject who answered only on the basis of the number of pulleys and did not take account of how the pulleys were configured should answer these problems incorrectly.

Insert Figure 5 about here.
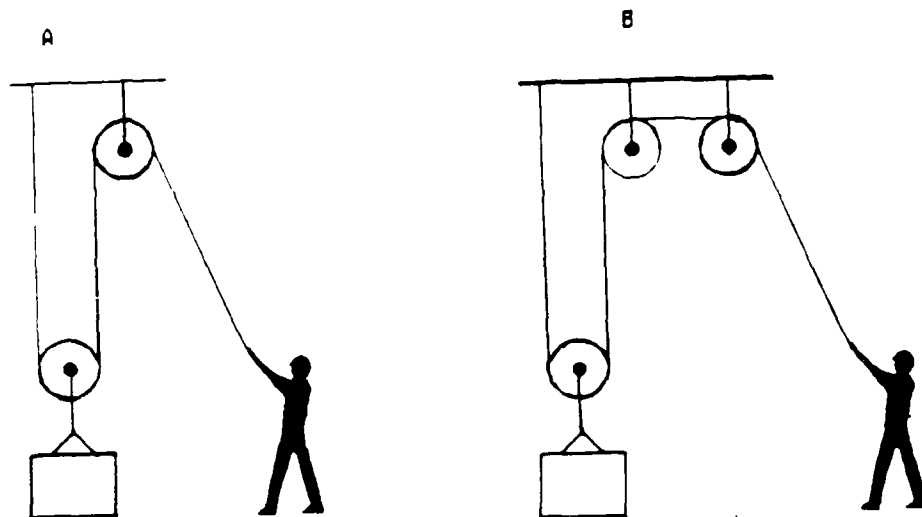

The third type of problem depicted systems with different mechanical advantage which were being used to lift different weights. In four of these problems, the difference in mechanical advantage between the systems compensated exactly for the difference in the weights to be lifted, so that the same effort was required in the two systems. In another four problems the difference in mechanical advantage was greater than the difference in weight, and in the remaining four problems of this type the difference in weight was greater than the difference in mechanical advantage. Only a subject who could quantify the mechanical advantage of a pulley system exactly would solve all of these problems correctly. A subject who had a preference for the weight attribute would solve only the second set of problems correctly while a subject with a preference for system attributes over weight would solve only the third set of problems correctly.

Subjects. The subjects were 27 undergraduate students at Carnegie-Mellon University. Ten of the students had taken two or more courses in physics at college level while the remainder had taken no college level physics courses.

Procedure. The subjects were tested individually and all were asked to give verbal protocols while solving the problems. Some subjects gave concurrent protocols as requested, while others (who constituted a majority) did not comply and were prompted to give retrospective protocols after solving each problem.

Comparisons will be made between high-scoring and low-scoring subjects. For consistency with Experiment 1, the high-scoring and low-scoring groups were defined by a discontinuity in the distribution of overall scores. Twelve subjects, all of whom solved at least 34 of the 44 problems correctly, were assigned to the high-scoring group. The remaining fifteen subjects, assigned to the low-scoring group, all solved 31 or fewer problems correctly. Seven of the twelve subjects in the high-scoring group and four of the fifteen subjects in the low-scoring group had studied physics at college level.

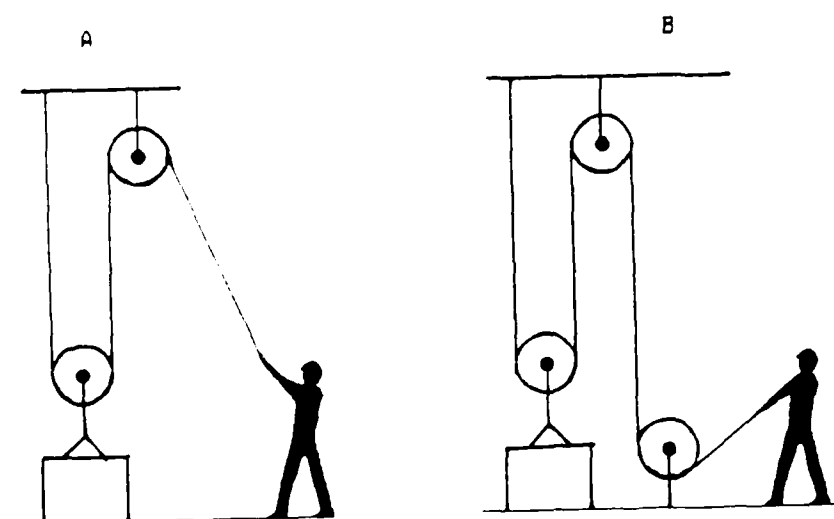Figure 5: Problems involving an irrelevant pulley

Example 6:

With which pulley system does the man have to pull with more force to lift the weight?

    A
    B
    If no difference.
    mark C.

Example 7:

With which pulley system does the man have to pull with more force to lift the weight?

    A
    B
    If no difference.
    mark C.

Results and Discussion

The general account of how subjects solved the test items, proposed on the basis of Experiment 1, also characterized the performance of subjects in Experiment 2. However the results of Experiment 2 allowed a greater range of solution processes to be identified, which produced a more precise measurement of the sources of individual differences in performance. As before, we first consider some solution processes that were general to the entire group of subjects and later focus on the sources of individual differences in performance.

The repertoire of rules used in Experiment 2 was inferred from the subjects' protocols and response patterns, which were generally in agreement. Subjects were classified as using a rule when their explanations or responses were consistent with that rule on at least 3 out of 4 problems of a particular type. Classifications for each of the 27 subjects were made on the basis of seven problem types, which correspond to the problem types listed in Table 7 (with the exception that only one classification was made on the basis of the three types of problems in which mechanical advantage and weight were varied). Separate classifications made on the basis of explanations and response patterns agreed in 151 (79.9%) of the 189 instances. In a further 29 (15.3%) instances the protocol data allowed subjects' responses to be classified when their response patterns seemed inconsistent with any rule. For example, some subjects used the rule that a system with a greater mechanical advantage requires less effort, but computed mechanical advantage incorrectly, so that the rule could not be inferred from their response patterns. Other subjects switched from one rule to another in the course of the experiment, a switch obvious from the protocols.

The repertoire of rules observed in this experiment was larger than in Experiment 1, reflecting the fact that protocols were collected from 27 subjects, as compared to 5 in Experiment 1. We can ask how the number of rules increases as we sample more subjects. To answer this question we took 10 random samples of sizes 1 to 15 from the sample of 27 subjects and assessed the average number of rules used by different sized samples. The average number of rules used by one subject was 3.9, and the number of rules increased by about 1.2 for each additional subject as the sample size increased from 1 to 8. Beyond this point, the number of rules increased negligibly with additional subjects. This analysis suggests that the number of common rules in this task is about 13 and that all the rules can be observed by sampling a relatively small number of subjects. As Table 8 shows, the rules most commonly used in Experiment 2 were also used in Experiment 1.

Insert Table 8 about here.

Individual Differences. The results of Experiment 2 suggested a more precise characterization of individual differences in mechanical ability. We added to our understanding of the three abilities underlying performance, observed in Experiment 1, by observing how subjects treat irrelevant information when it conflicts with relevant information, observing the types of rules used by consistent and less consistent subjects and observing how subjects treat conflicting relevant information. In the case of each of these sources of individual differences, we examine the range of solution processes observed in Experiment 2, the relation of these processes to accuracy in solving the pulley problems, and the extensions to the simulation model necessary to characterize these solution processes.

Ability to Identify Relevant Attributes. Two contrasting hypotheses were outlined

Table 8: Rules used by the Subjects in Experiment 2.

Rule	Number of Subjects who used the Rule.
A system with ... requires less force:	
less weight*	27
more pulleys*	18
larger pulleys*	13
more rope - weight attachments	8
greater mechanical advantage	8
less weight/mechanical advantage	7
less height*	6
more height	5
more load-bearing ropes*	5
less weight per load-bearing rope	3
less weight per pulley*	3
smaller pulleys	3
more rope-ceiling attachments	2
less pulleys	1
more movable pulleys*	1
less rope-pulley attachments	1

* Rules that were also observed in the protocols in Experiment 1.

concerning the treatment of irrelevant attributes by low-scoring subjects in Experiment 1. One hypothesis was that low-scoring subjects consider an irrelevant dimension relevant only if there are no other distinguishing attributes. The opposing hypothesis was that low ability subjects think that irrelevant attributes are relevant regardless of any other variation. The results indicated that some subjects in Experiment 2 are best described by the first hypothesis while other subjects are best described by the second hypothesis. Ten subjects, all low-scoring, considered the irrelevant attribute, height, to be relevant. Five of these subjects showed a consistent preference for rules based on relevant attributes when relevant and irrelevant attributes were covaried, indicating that they can be described by the first hypothesis. The remaining 5 subjects either showed no preference for the relevant attribute or preferred the irrelevant attribute, indicating that they can be described by the second hypothesis.

The results demonstrate that it is possible to order subjects with respect to the way they treated irrelevant information. For a particular irrelevant attribute, some subjects showed no preference for rules based on relevant attributes over rules based on this attribute, other subjects based their comparisons on the irrelevant attribute only if relevant attributes were not varied, and still other subjects understood the attribute to be irrelevant. This ordering of subjects was related to performance on problems in which both relevant and irrelevant attributes were varied ($r = .86$). Subjects who showed no preference for rules based on relevant attributes had a lower proportion of correct responses (.65) on these 16 problems than subjects who preferred relevant attributes (.74), which was not statistically significant. As expected, subjects who understood the attributes to be irrelevant solved a significantly greater proportion (.98) of problems of this type correctly ($t(11) = 6.7$, $p < .001$, as indicated by a two-sample t-test). The classification of subjects was also related to total performance on the test ($r = .82$). It is possible that these three types of subjects that we identified in the context of understanding of pulley systems represent different stages in the development of understanding of mechanical systems in general as a person gains more experience with these systems.

Preferences for relevant attributes over irrelevant attributes can be simulated in the existing framework of the model. However the model would have to be extended to account for a strategy used by some subjects who took both relevant and irrelevant attributes into account. When both a relevant and an irrelevant attribute were varied in a problem, these subjects did not make a random choice between the answers predicted by their conflicting rules (as would the present implementation of the model) but instead tried to assess the size of the effect of each attribute on the effort. This strategy, which we call the compensation strategy, will be discussed further below.

Consistency of Rule Use. Examination of subjects' consistency in solving problems that varied mechanical advantage revealed that the most consistent subjects compared the pulley systems directly on the basis of their relative mechanical advantage rather than on the basis of visible attributes that are correlated with mechanical advantage. The basis of the consistency was that once these high-scoring subjects computed the mechanical advantage of a particular pulley system in the earlier problems, they then retrieved the mechanical advantage for this system when they encountered it again on later problems. Subjects used one of two approaches to compute the mechanical advantage of pulley systems, both of which took into account how the components of the pulley system were configured. Some subjects computed the mechanical advantage by analyzing the balance of forces in the system, while others retrieved the mechanical advantage of simpler systems from memory and computed the advantage of other systems by matching their features with features of a simpler system whose mechanical advantage they knew. The latter strategy produced many errors, especially in computing the mechanical advantage of systems

(3) and (4). shown in Figure 6. For example. a subject might erroneously say that system (3) has a mechanical advantage of 2 because it is essentially the same as system (2). the middle rope erroneously being judged as irrelevant.

Insert Figure 6 about here.

Subjects were classified as quantitative if they computed the the mechanical advantages of the two systems and related the effort directly to the mechanical advantage. qualitative if they explained their answers using rules based on relevant attributes. and inconsistent otherwise. Quantitative subjects used the same rule. based on the mechanical advantage of the system. on all problems. Qualitative subjects typically explained their answers consistently in terms of one rule when solving problems in which several rules converged on the correct answer but were less consistent in justifying their answers to problems in which their rules dictated different answers. Inconsistent subjects did not give consistent explanations or answers. even in problems where several rules converged on the correct answer. This classification was related to total performance on the test ($r = .74$). Eight high-scoring and one low-scoring subject were classified as quantitative while four high-scoring and ten low-scoring subjects were classified as qualitative. Four subjects. all of whom were low-scoring. were inconsistent in their use of rules.
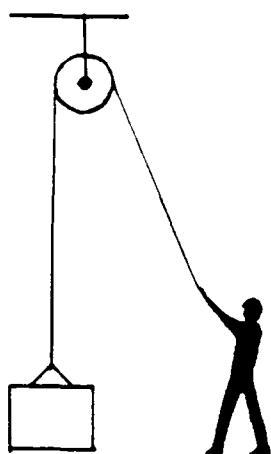
Quantifying the mechanical advantage of a pulley system depends on the ability to understand configural properties of the system. Thus. subjects who computed mechanical advantage were better able to recognize a pulley that was irrelevant to the mechanical advantage of a system because of the way it was configured. All 9 subjects who computed mechanical advantage recognized an irrelevant pulley as such if it was attached to the ceiling. and 3 of these subjects recognized an irrelevant pulley attached to either the ceiling or floor. Only 6 of the 14 subjects who used qualitative rules recognized an irrelevant pulley attached to the ceiling and one of these recognized an irrelevant pulley attached to either the ceiling or the floor. The ability to use rules consistently was therefore highly correlated (.66) with the ability to recognize an irrelevant pulley. The recognition of irrelevant pulleys by subjects who computed mechanical advantage is further evidence that these subjects understood mechanical advantage to depend. not just on what components a pulley system contains. but also on how these components are configured.

While quantitative subjects tended to have higher overall scores on the test. they did not score significantly higher than qualitative subjects on problems that varied only mechanical advantage. These two groups of subjects made errors for different reasons: the quantitative subjects' errors arose from inaccuracies in computing mechanical advantage while the qualitative subjects' errors arose from the use of rules that gave incorrect answers to some of the problems.
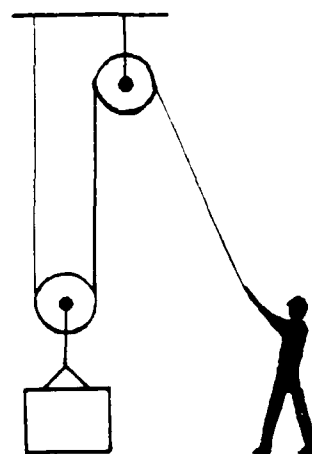
A number of extensions to the existing simulation model would account for the ability of subjects to compare pulley systems directly on the basis of mechanical advantage. The simulation for some subjects would have to include productions that can compute a value for the mechanical advantage of a pulley system by analyzing the balance of forces in the system. Moreover. the simulation model could be given the capacity to store the computed value of the mechanical advantage of particular pulley systems. The pulley systems in a new problem could then be matched against these stored representations and the mechanical advantage either retrieved or estimated on the basis of the number of shared attributes of the new pulley system and the retrieved representation.

Ability to combine two relevant attributes. In problems that required the quantitative
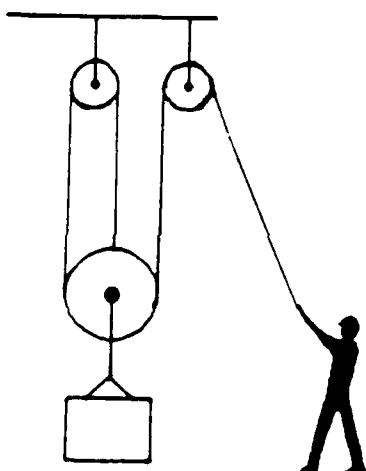
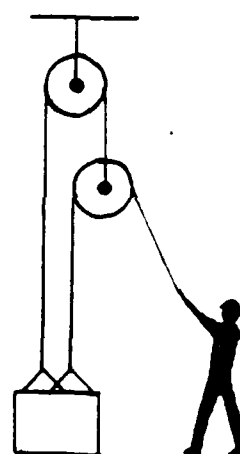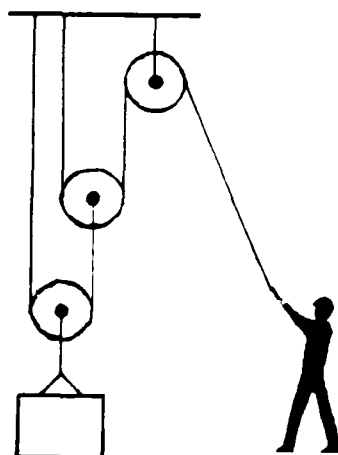Figure 6:   The five pulley systems depicted in the problems.

(1) M.A. = 1

(2) M.A. = 2

(3) M.A. = 3

(4) M.A. = 3

(5) M.A. = 4

combination of two relevant attributes (weight and mechanical advantage), subjects used one of the three strategies observed in the protocols in Experiment 1 to solve the problem. One strategy was to compute the effort directly by computing a ratio of the weight to some attribute of the system, such as the mechanical advantage or the number of pulleys. The second strategy was to use a principle whereby differences in mechanical advantage are considered to compensate for differences in weight. The third strategy was to use only one of the applicable rules, selected on the basis of a preference ordering.

Subjects who used the ratio strategy were of two types. The first type (2 subjects) performed like the high-scoring subject whose answers were simulated in our model. These subjects compared the pulley systems using qualitative rules based on relevant attributes such as the number of pulleys and computed the efforts involved only when the problem could not be solved using this easier comparison. The values that they computed for the efforts were incorrect because they were based on relevant attributes that are not correct indicators of the mechanical advantage of a pulley system. The other type (9 subjects) computed the mechanical advantage of the pulley systems in all problems and determined the efforts by computing the ratio of the weights to the mechanical advantage of the pulley systems. These subjects used a quantitative approach to solving all problems.

The majority of subjects who used rules based on relevant attributes used the compensation strategy to solve problems in which both weight and mechanical advantage were varied. Subjects who used this strategy tried to estimate the size of the difference between the two pulley systems on the two attributes that were in conflict. i.e.. weight and mechanical advantage. On the basis of these estimates they decided that either the difference in one attribute outweighed the difference in the other attribute or that the two differences compensated for each other. Compensation was similar to computing ratios in the respect that subjects who used this strategy understood that one attribute (e.g.. number of pulleys) could compensate for another (e.g.. weight). However it was dissimilar in the respect that it did not involve exact quantification of the effects of the two attributes.

The ability to combine information about two relevant attributes (weight and mechanical advantage) was related to total performance on the test ($r = .70$). Of the 12 subjects who combined these attributes quantitatively. 9 were high-scoring and 3 were low-scoring. The compensation strategy was used by the remaining 3 high-scoring subjects and 5 of the low-scoring subjects. The other low-scoring subjects either had a preference for the weight attribute (3). were inconsistent (2). or did not experience conflict in these problems because their rules stated that systems that actually had greater mechanical advantage required more effort (2).

The strategies that subjects used in problems varying mechanical advantage and weight were reflected in the patterns of performance on these problems. We would expect subjects who computed ratios to have similar scores on all the problem types. This was not the case but. as shown in Figure 7. subjects who used the ratio strategy did show a smaller range in performance (2.2 to 3.3 problems correct) than subjects who used the compensation strategy (0.5 to 2.5 problems correct) and subjects who preferred an answer based on weight (0.0 - 3.1 problems correct). Subjects using the ratio strategy made most of their errors in computing mechanical advantage. Subjects using the compensation strategy could only estimate the size of the difference in mechanical advantage and therefore they had much lower levels of performance on problems in which the weight difference compensated exactly for the difference in mechanical advantage than in the other problem subtypes. As expected. the pattern of responses for subjects who had a preference for the weight attribute was a high level of performance in problems in which the weight difference was greater than the difference in mechanical advantage and a very low level in the other

sub-types of problems in this category (in which mechanical advantage and weight vary).

The combinatorial rules that simulated the compensation strategy used by the low-scoring subjects in Experiment 1 would have to be extended to account for the fact that some subjects who used this strategy took the size of the difference into account. The determination of what was a large and what was a small difference depended on which attributes of pulley systems subjects considered relevant to their function and thus would have to be coded separately for each subject who used the compensation strategy.

Insert Figure 7 about here.

Relation of Specific Abilities to Total Performance. More precise information was available from Experiment 2 concerning the range of individual differences in each of the factors identified in Experiment 1. Subjects were assigned scores corresponding to each of the abilities as follows. For ability to identify relevant attributes, they were given a score of 2 if they never considered an irrelevant attribute to be relevant. 1 if they preferred relevant to irrelevant attributes, and 0 if they had no preference or preferred irrelevant to relevant attributes. For consistency, subjects were given a score of 2 if they used a quantitative rule consistently. 1 if they used qualitative rules consistently and 0 if they were inconsistent. For ability to quantitatively combine information, they were given a score of 2 if they computed ratios. 1 if they used the compensation strategy and 0 otherwise. Each of the ability scores had a correlation with the overall score that lay between .70 and .82. Together they accounted for 81.5% of the variance in overall performance. This figure, compared to 38.6% in Experiment 1 indicates a considerable improvement in accounting for the total score from the more precise measures of individual differences in Experiment 2. A possible fourth ability identified in Experiment 2. was the ability to take the configuration of the pulley system into account in computing mechanical advantage. This ability was highly correlated with consistency of rule use $(r = .66)$ and did not add to the variance accounted for by the other three abilities.
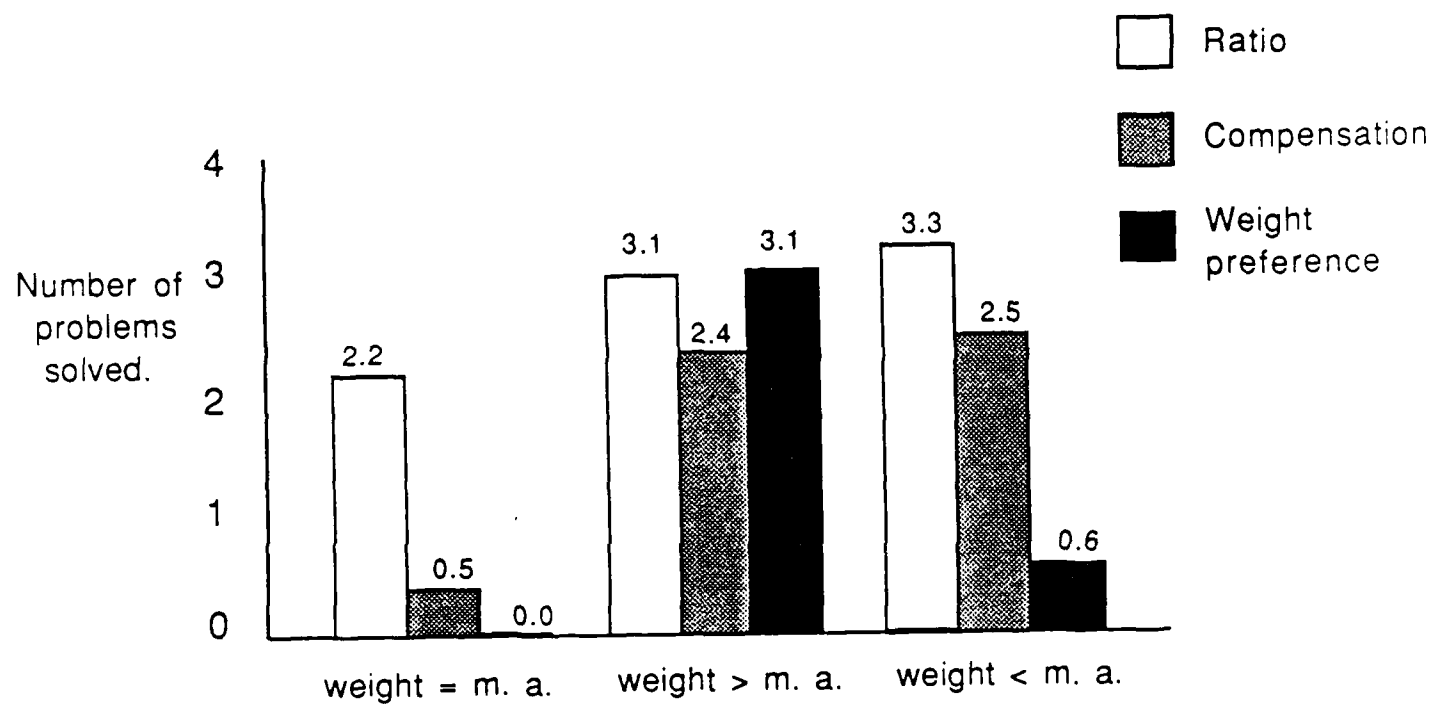
General Discussion.

Summary of results.

The research reported in this paper provides both a general model of the processes involved in solving items from tests of mechanical ability and identifies sources of individual difference in performance on these tasks. It was found that subjects encoded mechanical systems in terms of attributes of systems that they considered relevant to their function. These attributes could be surface, visible features (e.g., the number of pulleys in a system) or abstract properties (e.g., the number of units of force pulling up on a weight). They could be attributes of the whole system (e.g. its mechanical advantage) or properties of components (e.g. the size of a pulley). Comparison of the pulley systems by different subjects was based on rules that expressed a relation between one or more of these attributes and the attribute in question. i.e., the effort required to lift the weight with the pulley system.

Figure 8 presents a schematic description of the range of individual differences that we found in people's ability to solve pulley problems. According to the description of individual differences presented in the figure. low-scoring subjects are characterized as using rules based on visible components of pulley systems. These rules are qualitative, the attributes on which they are based can be either relevant or irrelevant. and subjects have no clear preferences among their rules so that their responses appear inconsistent with any particular rule. High-scoring subjects. on the other hand. have rules that are quantitative

Figure 7: Performance in Experiment 2 on problems varying mechanical advantage and Weight.

and take configural properties of the system into account. They prefer rules based on attributes that are highly correlated with mechanical advantage. Although we make no strong claims about the exact location of the acquisition of specific abilities along the continuum, the point at which the label for each ability is placed in the figure corresponds to the proportion of subjects in Experiment 2 whose performance demonstrated possession of this ability. For example, if 50% of subjects demonstrated that they had a particular ability, it would be placed half way between the low ability end and the high ability end of the continuum.

Insert Figure 8 about here.

The experiments suggested three abilities that account for the range of individual differences described in Figure 8, namely ability to differentiate relevant from irrelevant attributes, ability to use rules consistently, and ability to quantitatively combine information about two relevant attributes. A simulation model specified mechanisms that can account for these three sources of individual differences. The model successfully simulated the performance of a high-scoring and a low-scoring subject indicating the sufficiency of the theoretical proposal. The model suggested that the process of applying rules is similar for high-scoring and low-scoring subjects, but that the content of the rules changes with increases in mechanical ability.
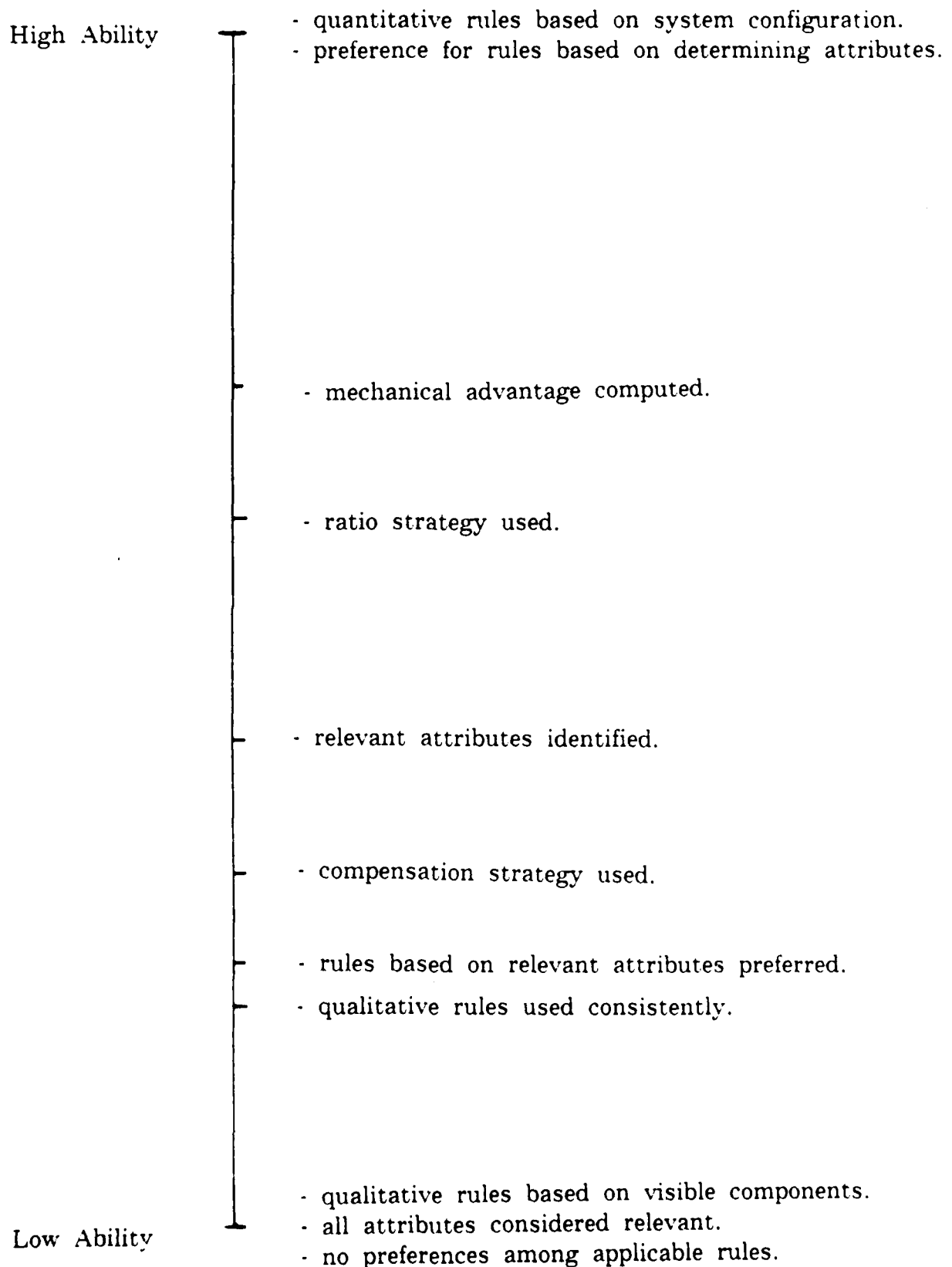
Mental models of pulley systems.

We have characterized performance at different levels of mechanical ability in terms of rules which relate the function of a mechanical system to attributes of the systems components and their interactions. *Rules of this type imply a causal connection between* some attribute of a pulley system and the amount by which a system magnifies an input force. Thus the rules that characterize performance of an individual on our pulley problems reflect his understanding of the causal interrelations among attributes of pulley systems. This causal understanding of a mechanical system constitutes a mental model of the system.

A key difference between the mental models of high-scoring and low-scoring subjects is that low-scoring subjects used only qualitative models while high-scoring subjects used both qualitative and quantitative models. This is demonstrated by our subjects' strategies for combining information about two or more relevant attributes when these attributes suggested conflicting answers. In this situation, low-scoring subjects evoked qualitative models while high-scoring subjects evoked quantitative models. In a qualitative model, attributes of pulley systems are coded in terms of a qualitative comparison with corresponding attributes of other pulley systems. A qualitative model also includes rules *relating attributes of mechanical systems, situations in which it is appropriate to apply* these rules, and preferences among these rules. Preferences can resolve conflicts between rules that are equally applicable in a given situation, but there is no simple way in a qualitative model to resolve conflicts between rules with equal preference. Our results thus provide converging evidence that qualitative models precede quantitative models (Forbus & Gentner, 1986; White & Frederikson, 1986).

It is interesting that some subjects who used a qualitative approach attempted to compare the size of the difference between the two systems for each of the conflicting attributes. They were thus going beyond an ordinal scale of measurement which was characteristic of the qualitative approach. This may be an important step in the

Figure 8: Schematic representation of the progression from low to high ability.

High Ability
- quantitative rules based on system configuration.
- preference for rules based on determining attributes.

- mechanical advantage computed.

- ratio strategy used.

- relevant attributes identified.

- compensation strategy used.

- rules based on relevant attributes preferred.
- qualitative rules used consistently.

Low Ability
- qualitative rules based on visible components.
- all attributes considered relevant.
- no preferences among applicable rules.

progression from a qualitative to a quantitative understanding of pulley systems. which is demonstrated by the fact that one subject made such a progression in the course of the experiment.

A central issue in the literature on mental models of physical systems has been how to characterize naive intuitions of physical systems (intuitions based on experience in the physical world rather than formal instruction in physics). One approach has been to characterize the intuitions in terms of Qualitative Process Theory (Forbus, 1983). In Qualitative Process Theory. quantities are defined. not by absolute values. but in terms of their comparison with other quantities. Functional relationships. similar to the rules in our model. express qualitative proportionality between different quantities. Determining the influence of various processes on a quantity involves determining whether various processes will increase or decrease the quantity. However. the relative magnitude of the increasing or decreasing influences is not always known. as was shown by some of our subjects who used the compensation approach. and were unable to compare the relative effects of mechanical advantage and weight on the amount of effort required to lift a load. Our data provide support for Qualitative Process Theory as a characterization of naive intuitions about physical systems.

Expert performance appears to include both the quantitative reasoning that distinguishes experts. as well as qualitative. causal reasoning of the type demonstrated by novices (De Kleer. 1985: diSessa. 1983). Our results suggest that high-scoring subjects are flexible problem solvers who can use either qualitative or quantitative mental models. depending on the demands of the problem. Some high-scoring subjects in our study did not resort to the use of a quantitative model unless the problem required it. If the relative effort required to lift the weight in the two pulley systems could be determined by comparing a single attribute (rather than by actually computing the two efforts). then just that attribute was evaluated and compared. For these subjects. qualitative models. which may be simpler to use. were invoked when they were adequate to the task. Thus. there may be a least-effort principle operating.

Mental Models of Low-Scoring Subjects. Although our low-scoring subjects had little if any experience with pulley systems before taking part in our experiments. they clearly demonstrated some consistent conceptions of the behavior of pulley systems. For example. all subjects understood that a pulley system that was lifting a heavier weight required more effort. and in problems where only one attribute distinguished the depicted pulley systems. the responses of the majority of low-scoring subjects could be classified as rule governed (see Table 3). Thus. although the responses of low-scoring subjects were less consistent than the responses of high-scoring subjects. this inconsistency arose partly from an inability to resolve conflicts between multiple rules. rather than from an absence of rules about the behavior of pulley systems.

The consistency of naive mental models has also been an issue in previous literature. In this literature. naive mental models have sometimes been characterized as a systematic body of knowledge that is consistent across individuals (McCloskey, 1983) and sometimes as a collection of knowledge fragments that are brought to bear on different problems (diSessa. 1983). Our characterization of knowledge as a set of rules relating attributes of mechanical systems to their function suggests that mental models consist of elements of knowledge. and that consistency arises from a precise specification of the conditions of application of rules and consistent preferences among rules.

Low-scoring subjects' mental models of pulley systems show some typical naive intuitions about physical systems. For example. one common misconception that was first

recognized as such by Gallileo (Einstein & Infeld, 1938), is that forces dissipate and that a constant force is required to keep a body in motion (Clement, 1983: diSessa, 1983; White, 1983). This misconception would lead to the incorrect prediction, made by some of our low-scoring subjects, that higher pulley systems, with a greater length of rope between the point at which the force is exerted and the point at which the rope is attached to the weight, would require more force to lift a given weight. This misconception is probably due to the experience of living in a world with friction, as noted by White (1983).

Low-scoring subjects' misconceptions of pulley systems may also reflect their use of inappropriate analogies to more familiar mechanical systems. Distance is a critical attribute in the operation of a lever, so a person might draw an incorrect analogy from the lever to the pulley and think that if the pulley system is higher (and therefore includes a longer rope) a greater magnification of force results. Some of our subjects gave answers consistent with this analogy and stated that the man pulling the longer rope was getting more leverage. A similar analogy might account for the responses of subjects who answered that a pulley system with larger pulleys would require less effort. In this case the analogy might be to gear systems in which gear size is a critical variable. Analogies of this type, based on literal similarity between objects in the two situations being compared, are highly accessible and are typical of novices (Gentner, 1983; Forbus & Gentner, 1986).

Individual differences and development.

A striking feature of the range of solution processes used by subjects with different mechanical ability is their similarity to the developmental stages observed by Siegler (1978, 1981) in his analysis of young children's understanding of a balance beam. We found that there were subjects who always chose the system with the greater weight when weight and mechanical advantage were in conflict and considered attributes of the pulley system only when the weights of the two systems were equal. These correspond to children at stage 2 in Siegler's account, who consider the distance of weights from the fulcrum only when these weights are equal. Our results also suggest, in parallel with Siegler's, that the idea of compensation precedes the ability to quantitatively combine two attributes. The subjects in our study who used the compensation strategy correspond to Siegler's stage 3 children. Children at this stage always consider both weight and distance, but when one side of the balance beam has more weight and the other has its weights at a greater distance, they have no consistent formula for resolving the conflict.

The parallel between our findings and developmental findings such as Siegler's suggests the intriguing hypothesis that the processes that underlie the development of mechanical abilities also characterize differences along an individual difference dimension. Individual differences are sometimes thought of as static abilities, whereas developmental differences are typically thought of as changing with age and experience. Comparison of our results with the results of developmental studies suggests that differences in abilities at different points along an individual difference continuum are very similar to changes in abilities that occur with development. A further analogy can be drawn to the progression that occurs in the historical evolution of scientific understanding of a domain, for example, the history of understanding of the pulley system. Similar analogies have been made by Carey (1985) in her analysis of the development of understanding of biological concepts. Her research suggests that concepts develop with increasing knowledge of a domain. This knowledge can occur as a child gains more familiarity with the natural world, as an adult becomes an expert in a particular domain, or as scientists gather data about some natural phenomenon.

Viewing individual differences in mechanical ability as a progression of mental models that develop with experience is consistent with conclusions from the psychometric literature, namely that mechanical ability is a measure of understanding acquired through general exposure to tools and machinery (Cronbach, 1984). Although our study provides no longitudinal data, our characterization of performance at different levels of mechanical ability provides a framework within which to speculate about what mechanisms might underlie the progression from low to high ability with exposure to machinery and to formal instruction in physics.

Progressions of Mental Models. A theory of the progression from low to high ability has to account for advances along a number of different dimensions (see Figure 8). One advance is from having no consistent preferences among rules based on different attributes to preferring rules based on attributes that are highly correlated with mechanical advantage. A strengthening mechanism that increases the preference of successful rules and decreases the preference of unsuccessful rules would produce a gradual increase in the relative strength of more correct rules over less correct rules as a person gains more experience with mechanical systems. A second advance is from encoding the systems in terms of basic physical components (e.g., pulleys and rope strands), to recognizing larger patterns in systems that involve configurational relations among a number of system components (e.g., a rope going over a pulley that is attached to the ceiling.) Another learning mechanism, chunking, (Rosenbloom & Newell, 1987) can account for this type of organization of knowledge of a domain by forming and storing patterns (chunks) that are structured collections of more elementary patterns that were present at an earlier stage of expertise. A third advance is the progression from a qualitative to a quantitative model of mechanical advantage that enables the subject to quantify the extent to which mechanical advantage can reduce the effort required to lift a particular weight. It is likely that the abstract concepts of force and mechanical advantage are a result of formal instruction, except in very rare cases of extremely high mechanical ability. Our performance model provides a theoretical framework within which these advances can be encompassed.

Generalization to reasoning in other domains.

Reasoning in many other domains requires the processes measured by psychometric tests of mechanical ability, i.e., deciding which attributes are relevant to a judgment being made, how these attributes are related to the attribute to be judged, and how the information about these attributes can be combined to form an overall evaluation of each alternative. For example, when diagnosing an illness, a person must decide which environmental factors, internal factors, and symptoms are present and how these are related in various illnesses, and combine this information to diagnose which illness is present. Similarly, when choosing consumer goods, a person often has to evaluate two or more alternatives that differ from each other in a number of attributes, for example price and various indices of quality.

Our research suggests that people decide which attributes of a mechanical system are relevant to judging mechanical advantage on the basis of causal models of mechanical systems which they form as a result of experience with these systems. Similar conclusions have been made in studies of naive medical understanding. For example, Meyer, Leventhal and Gutman (1985) suggested that people interpret illnesses in the framework of common sense models that they form as a result of the illnesses they are exposed to. These models are similar to mental models of mechanical systems in that they emphasize causal interconnections, in this case, between environmental factors, the susceptibility of the body to different diseases, and physical symptoms. Like the mental models of mechanical systems that we identified, mental models of illness have been found to differ for people with different amounts of knowledge of illness (Taylor, 1982). The interpretation of situations in

terms of causal models may be a general characteristic of reasoning in domains that have underlying principles of operation.

Once a person has decided which attributes of a situation are relevant to making a judgment, she has to combine the information from these attributes to form an overall evaluation for each alternative. The processes that our subjects used in dealing with multiple attributes may also be involved in other forms of judgment. For example, Slovic and Lichtenstein (1968) have shown that people often give different weights to different attributes that they consider relevant to a judgment. These weights can be compared to the preferences demonstrated by our subjects for rules based on some attributes of pulley systems over rules based on other attributes. Studies in other domains have also shown that people do not always consider all the information available to them in making judgments, but instead base their judgments on an evaluation of a subset of the attributes that vary in the situation (Slovic, 1969; Wright, 1974). The behavior of our low-scoring subjects, who based their judgments on only one attribute when faced with problems in which two relevant dimensions provide conflicting cues, can be seen as an instance of this type of information reduction. Thus some of the processes used by our subjects in solving pulley problems may reflect more general heuristics for making judgments about alternatives that vary on a number of dimensions.

In conclusion, we have viewed mechanical reasoning as a process of applying inference rules that relate attributes of machines to their function. This approach has allowed us to distinguish between the process of applying rules and the content of the rules applied. The experimental results and the model have demonstrated that the processes of applying rules can be similar at different levels of ability. These processes may be shared with other reasoning tasks that involve evaluating alternatives varying on a number of different attributes. The experimental results also demonstrated that the content of the rules can vary with individual differences in mechanical ability. The content reflects people's causal models of machines, formed as a result of relevant experience. Mental models in other domains are likely to be governed by similar principles of operation.

References

Bechtoldt, H. P. (1972). Review of the Bennett Mechanical Comprehension Test. In Buros, O. K. (Ed.), *Seventh mental measurement yearbook*. Highland Park N.J.: Gryphon Press.

Bennett, C. K. (1969). *Bennett mechanical comprehension test*. New York: The Psychological Corporation.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Clement, J. (1983). A conceptual model discussed by Galileo and used intuitively by physics students. In Gentner D. & Stevens, A. L. (Ed.), *Mental models*. Hillsdale, N.J: Erlbaum.

Cox, J. W. (1928). *Mechanical aptitude*. London: Methuen.

Cronbach, L. J. (1984). *Essentials of psychological testing (4th edition)*. New York: Harper and Row.

De Kleer, J. (1985). How circuits work. In D. G. Bobrow (Ed.), *Qualitative reasoning about physical systems*. Cambridge MA: The MIT Press.

diSessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. L. Stevens (Ed.), *Mental models*. Hillsdale, N.J: Erlbaum.

Einstein, A. & Infeld, L. (1938). *The evolution of physics*. New York: Simon and Schuster.

Forbus, K. D. (May 1983). *Qualitative Process Theory* (A.I. Memo 664). Massachusetts Institute of Technology.

Forbus, K. D. & Gentner, D. (December 1986). *Learning physical domains: Toward a theoretical framework* (Tech. Rep.). University of Illinois at Urbana-Champaign.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy . *Cognitive Science, 7*, 155-170.

Ghiselli, E. E. (1955). The measurement of occupational aptitude. *University of California Publications in Psychology, 8*(2), 101-216.

Laird. J. E., Newell. A & Rosenbloom. P. S. (in press). Soar: An architecture for general intelligence. *Artificial Intelligence.* .

McCloskey. M. (1983). Naive Theories of Motion. In D. Gentner & A. L. Stevens (Ed.). *Mental models.* Hillsdale NJ: Erlbaum.

Meyer. D.. Leventhal. H., & Gutmann. M. (1985). Common-sense models of illness: The example of hypertension. *Health Psychology.* 4(2). 115-135.

Rosenbloom. P.S.. & Newell. A. (1987). Learning by chunking: A production-system model of practice. In D. Klahr. P. Langley. & R. Neches (Ed.). *Self-modifying production system models of learning and development.* Cambridge MA: Bradford Books/ The MIT Press.

Siegler. R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.). *Children's thinking: What develops?.* Hillsdale. NJ: Erlbaum.

Siegler. R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development.* 46(2). 1-149.

Slovic. P. (1969). Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology.* 53(4). 255-263.

Slovic P. & Lichtenstein. S. (1968). Relative importance of probabilities and payoffs in risk taking. . *Journal of Experimental Psychology.* 78(3). 646-654.

Smith. I. M. (1964). *Spatial ability: Its educational and social significance.* London: University of London Press.

Stenquist. J. L. (1922). *Mechanical aptitude tests.* Yonkers: World Book Co.

Taylor. S. E. (1982). Social cognition and health. *Personality and social psychology.* 8(3). 549-562.

Vernon P. E. & Parry. J. B. (1949). *Personnel selection in the british forces.* London: University of London Press.

White. B. Y. (1983). Sources of difficulty in understanding newtonian dynamics. *Cognitive Science.* 7. 41-65.

White. B. Y. & Frederiksen. J. R. (March 1986). *Qualitative Models and Intelligent Learning Environments* (Tech. Rep.). BBN Laboratories.

Wright. P. (1974). The harassed decision maker: Time pressures. distractions. and the use of evidence. *Journal of Applied Psychology.* 59(5). 555-561.

Author Notes.

END

DATE
3-88
DTIC